THE SOCIETY OF THE SIGMA XI
DEVOTED TO THE
PROMOTION OF RESEARCH IN SCIENCE

National Lectureships
1939 and 1940

SCIENCE IN PROGRESS

SECOND SERIES

★ ★

Science in Progress

By

L. J. STADLER
F. W. WENT
J. F. FULTON
DOUGLAS JOHNSON
ALFRED C. LANE
H. P. ROBERTSON
CARL D. ANDERSON
DUNCAN A. MACINNES
J. W. BEAMS
J. C. HUNSAKER

Edited by

GEORGE A. BAITSELL

SECOND SERIES

New Haven
YALE UNIVERSITY PRESS

LONDON · HUMPHREY MILFORD · OXFORD UNIVERSITY PRESS

Printed in the United States of America

First Published, December, 1940
Second Printing, February, 1945

Preface

IN the Preface to the first volume of *Science in Progress*, published early in 1939, the hope was expressed that the distribution would warrant similar publication of the lectures presented in the 1939 and 1940 National Sigma Xi Lectureships. This has proved to be the case, and accordingly ten more notable contributions to scientific knowledge are now made available to scientist and layman in the present volume. It is expected that the third volume of the series, based on the 1941 and 1942 Lectureships, will be published in the fall of 1942.

The underlying purpose and the inherent values of these contributions by scientific leaders cannot be better expressed than they were by Dr. Harlow Shapley in the Foreword to the first volume.

"An encyclopedist of a century or so ago could gather into fruitful comprehension the facts and theories of all branches of science. A gifted individual was then able to know what conquests the techniques had achieved, into what fields of interpretation and imagination the current hypotheses were leading. But those days have gone. The capacity of the mind has not changed, while the activities and accomplishments of science have multiplied and become widely diverse. No longer is the botanist entirely conversant with the latest chemical theories, or the physiologist with gelactic structure. And it is a pity. If this unity were still possible, philosophy would be the richer and the sciences would gain by the wider understanding.

"The reader of the contributions to knowledge assembled in this book will find no synthesis of all the fields covered, but he will find, if attentive to detail, unification of another sort. He will note the employment of many of the same technical tools throughout the various branches of science and the general use of the most common instrument of all reasoned experimental

science—the balanced alternation of guiding hypothesis and experimental test.

"There is, however, another more subtle welding of the widespread sciences. For though we cannot hope to achieve, each for himself, a full and rounded view of the world of science, we can go far to envisage such a world and appreciate its variety. Whatever path the scientist chooses to travel, whether it be toward the minutest fragment of matter, or toward the wide stretch of the cosmos, or toward the still more mysterious activities of a human brain, he travels with one deep incentive—the satisfaction of intellectual curiosity, the understanding of some part of the world of nature.

"This, then, is the attainable unity that lies behind science and furnishes the common ground on which all its aspects may be approached. We can, by looking even briefly at the progress in diverse fields, obtain some vision of the new explorations and some understanding of the temper of the sciences. We can, perhaps, sense where they are going, what kind of a world they will reveal, how they affect both present and future human growth."

It is a great pleasure, as editor, to acknowledge the complete coöperation of the contributing authors and also of the publishers in the many details essential to the publication of this volume.

George A. Baitsell
National Secretary,
Society of the Sigma Xi

New Haven, Connecticut,
November, 1940

Contents

Illustrations

Science in Progress

I

THE EXPERIMENTAL ALTERATION OF HEREDITY

By L. J. STADLER

U. S. Department of Agriculture and University of Missouri

ALTHOUGH genetics is one of the youngest of the sciences, the alteration of heredity is an ancient problem. An early experiment is described in the Book of Genesis. Jacob tended the flocks of Laban, and received as his hire the streaked and spotted young. He insured the production of streaked and spotted offspring by the simple expedient of surrounding the breeding grounds with streaked rods. "And the flocks conceived at the sight of the rods, and the flocks brought forth streaked, speckled, and spotted."

This technic is no longer effective, although the same idea in various forms may be found among practical animal breeders to this day.

During the last century there was a great wave of interest in this problem. It centered upon the question of the inheritance of acquired characteristics. Is the dark skin of the negro the result of hereditary changes induced by intense sunburning through a period of generations? Is heredity modified by injury or disease? Is the improved speed of the trained race horse passed on to his descendants? Or the improved mind of the educated man?

Our grandfathers were deeply impressed with the importance of this problem, which Herbert Spencer declared was "the question which demands, beyond all other questions whatever, the attention of scientific men." A voluminous literature developed. The problem engaged the interest of the sociologist and the philosopher as well as the biologist. Much of the discussion was quite unrelated to experimental evidence.

But there were also many experiments. For example, it was found that guinea pigs could be made epileptic by partial section of the spinal cord. Epileptic individuals were found also among the progeny of the animals treated. Several investigators attempted, by cutting off the tails of mice at birth through a period of generations, to produce a tailless race. Similar tests were made of the inheritance of the most diverse effects of parental experience. In a considerable portion of these cases, the experiments were reported to be successful; other and similar experiments yielded negative results. A mass of evidence was accumulated from which the advocates of either side could prove their case.

Fifty years ago Weismann analyzed critically the whole body of evidence available at that time, and concluded that all of the positive results could be explained as errors in experiment or in interpretation. It was Weismann's contention that the germinal substance is entirely independent of the body, and that hereditary changes can come only from the alteration of this substance. So thoroughly did he demolish the evidence for the inheritance of acquirements, and so firmly did he establish the view that the germ plasm is completely insulated from the body, that there developed one of those curious scientific taboos that arise so often in the wake of a sweeping generalization. Many biologists came to regard the germ plasm as something apart and unreachable, so insulated from its environment that nothing that could be done to it could change it.

We know now that heredity can be altered by experimental treatment. The germinal substance is insulated from the body, as Weismann maintained, and there is no convincing evidence that any acquired characteristic has the slightest effect on heredity. But the germinal substance itself may be *directly* modified, so that hereditary changes result. These changes in heredity have been produced by various kinds of treatment. Some of the treatments are so effective that almost every germ cell of the treated individual is genetically altered in some way.

These experiments succeeded where the earlier ones had failed,

chiefly because in the interim something had been learned of the physical basis of heredity. The machinery of heredity was almost entirely unknown until the beginning of this century. We do not yet know it very thoroughly, but we have come a long way from the time when biologists could seriously attempt to breed a tailless race by cutting off tails.

The basic question is "What must be changed in the normal mouse to start a tailless race?" We are not looking merely for a mouse without a tail; the new mouse must transmit taillessness to his descendants. All that he can pass on must be included within the germ cell, for this is his only material connection with the next generation. Normal mice pass on materials which somehow determine normal mousehood, tail and all; tailless mice pass on materials somehow different, so that the tail does not develop. Tailless races are known; they are descended from individuals in which the germinal alteration occurred under unknown circumstances. Once established, the tailless condition is inherited with the same regularity as the normal; it has become the normal condition of the new race— as in fact it has in our own species. The problem of altering heredity is simply the problem of detecting these changes under controlled conditions, and then finding conditions under which they occur most frequently.

THE MATERIAL BASIS OF HEREDITY

The size of the male germ cell in man is shown in Figure 1, in comparison to the thickness of the paper on which this is printed. The sperm head alone is to be considered, since the rest is accessory equipment which is not concerned in heredity. This minute bit of substance carries within it everything which a human being inherits from his father, for literally he has no organic connection with his father except through this single cell.

What must we imagine to be contained within this microscopic drop of protoplasm? Hundreds of human traits have been shown to be inherited—characteristic forms of all sorts of organs of the body, characteristic mental and emotional traits, tendencies to

various organic defects and susceptibilities to various diseases. In addition there is the greater number of less definite traits which together make up the baffling family resemblances we see all around us.

All of these, if they have any material basis whatever, must be represented by materials within the single germ cell. These traits are not inherited as a unit, for the children of one family are not all the same. The hereditary traits are transmitted in all possible combinations with one another, and must therefore be due to separable elements which may occur in different combinations in the different germ cells of the same individual.

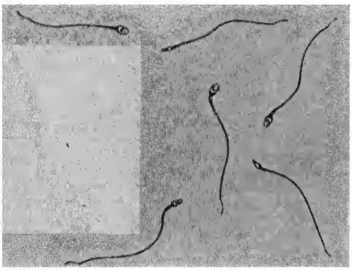


FIG. 1. Human sperm. Size of sperm is indicated by light colored area at the left, which gives approximate thickness of this page of paper at the same magnification.

We must therefore consider the germ cell to contain a very large number of specific constituents, each capable of producing distinctive effects. But in the development of the offspring this material must be subdivided among thousands of cells, and there is strong evidence that each of these receives regularly a precise duplicate of the particular combination present in the original cell. This may in turn be passed on to following generations for further subdivision.

This extraordinary degree of specificity, together with the capacity for unlimited subdivision, was long considered a hopeless bar to the explanation of heredity on any mechanical basis. What conceivable material could be so specific as to do all that heredity requires, and so unspecific as to permit unlimited subdivision? As recently as 1914 the eminent physiologist, J. S. Haldane, concluded: "The mechanistic theory of heredity is not merely unproven, it is impossible. It involves such absurdities that no in-

telligent person who has thoroughly realized its meaning and implications can continue to hold it."

Today biologists are agreed that heredity is explainable in mechanical terms. I cannot take time to consider in detail the evidence on which this conclusion is based, but will outline the mechanism as it is now understood.

The determiners of the separately inherited characteristics are not themselves visible, but they are carried in the chromosomes, which are microscopically visible constituents of the cell. The chromosomes of man, as they appear microscopically at a stage favorable for observation, are shown in Figure 2. There are 48 chromosomes in each body cell, and at every cell division each of the 48 is duplicated and divided, so that each of the two daughter cells resulting from division receives the full complement of 48 chromosomes. Accordingly each cell of the body contains 48 chromosomes, and these 48 chromosomes are duplicates of the 48 which were present in the fertilized egg.

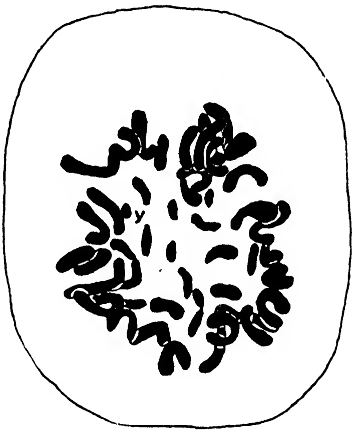


FIG. 2. Human chromosomes as seen in a male germ cell during maturation. (After Painter.)

Before fertilization the egg contains 24 chromosomes from the mother; at fertilization 24 chromosomes are added from the father. The 48 chromosomes present in each cell of the human body therefore include two sets of hereditary determiners, a full set from the mother and a full set from the father. In the production of germ cells by this individual, the chromosomes are again reduced to a single set of 24 in each sperm and each egg.

The identification of the individual determiners is much less

advanced in man than in certain lower animals and plants, since, obviously, man is not a very favorable object for experimental studies of heredity. But heredity in man, so far as it may be

FIG. 3. Chromosomes of *Drosophila melanogaster* (above) and gene maps of these chromosomes (below) as described on page 7. (After Bridges.)

studied by pedigree analysis, seems wholly analogous in mechanism to heredity in the more fully studied organisms.

The most thoroughly studied organism with respect to heredity is the fruit fly, Drosophila. This valuable insect, which almost seems to have been designed for the convenience of the

geneticist, produces three generations of offspring per month in the laboratory, has hundreds of children in each family, and lives on practically nothing at all. Instead of 24 pairs of chromosomes, as in man, Drosophila has 4 pairs. Probably more than 50 millions of these flies have been studied in genetic experiments during the past thirty years, and the pattern of their germ plasm has been thoroughly mapped. The Drosophila chromosome complex and gene loci maps are shown in Figure 3.

Each of the 4 chromosomes carries a long string of genes. How many genes there are in all we do not know, but several hundred of them have been identified and located, each in an exact position in one of the 4 gene strings. These genes are the individual determiners of heredity. Each of the detectable genes determines some inherited quality of the fly, in some cases a quality which produces a recognizable characteristic or several recognizable characteristics, in other cases a quality which may be detected only under special conditions or in the presence of the effects of several other genes.

For example, consider the gene *w*, which may be found on the map in Chromosome I near the left end. This was the first gene studied in the fruit fly. A fly was found with white eyes, instead of the red eye normally present. In crosses with the original type, the inheritance of the new characteristic was found to parallel the inheritance of the first chromosome. Later other characteristics dependent on the first chromosome were found, such as yellow body (*y*), cut wings (*ct*), and forked bristles (*f*). The relations of these genes to one another in inheritance provided the basis for their placement on the gene map.

It should be understood that the gene map is essentially a model, comparable with the atomic model of the physicist, rather than with the map of the geographer. It represents an assumed arrangement of hypothetical gene units, on the basis of which we may rationalize the results of experiments. No one has seen the gene for white eye, or analyzed it, or isolated it. We do not know what it is, or how it acts to produce the effect by which we recog-

nize its presence. But we do know that it is carried by this chromosome, that it is located well down toward the end, that its place in line is definitely beyond the gene *ec* and definitely before the gene *pr*. Whenever this precise portion of chromosome from a white-eyed race is present the characteristic effect is produced; when this chromosome region from a normal-eyed race is present the eye is normal. When both are present in the same fly, the eye is normal also; that is, the red-eye effect is dominant to the white-eye effect in the hybrid.

On the basis of this model the results of untried hybrid combinations may be predicted, just as the chemist may predict the results of the interaction of untried combinations of chemical substances.

In one respect this unpretentious model has a real advantage over the more elaborate atomic and molecular models of the physical sciences. It is subject to a direct visual check. The order of genes in Group I, as shown in Figure 3, was postulated solely on the basis of breeding experiments. It was obvious that this group of genes must be carried by the sex chromosome, a visually distinguishable chromosome. Further, certain breeding experiments showed which end of the gene string must correspond to the end of the chromosome which is attached to the spindle figure in cell division. The result is a definite correspondence of hypothetical gene map and visible chromosome. If portions of the visible chromosome could be broken off and lost or attached to other chromosomes, this would permit a crucial prediction of the parts of the gene string that should be affected. Precisely such chromosome changes may be experimentally produced by irradiation, as will appear presently, and the results have fully confirmed the reality of the gene maps.

The gene map of corn (maize) is shown in Figure 4. The corn plant has 10 chromosomes in each germ cell, and 10 pairs of chromosomes in each body cell. It is striking to see how closely the mechanism of heredity in this plant parallels that in a far removed animal organism, the fruit fly. The genes received by

a corn plant from its father and mother are very different from what a fruit fly receives, but the manner of transmission is in all essentials precisely the same.

If heredity is due entirely to the effects of genes carried by the chromosomes, the transmission of hereditary characteristics must be as simple and regular as the transmission of the chromosomes. If we consider any single hereditary quality, under conditions permitting its detection, it can easily be shown that this is true.

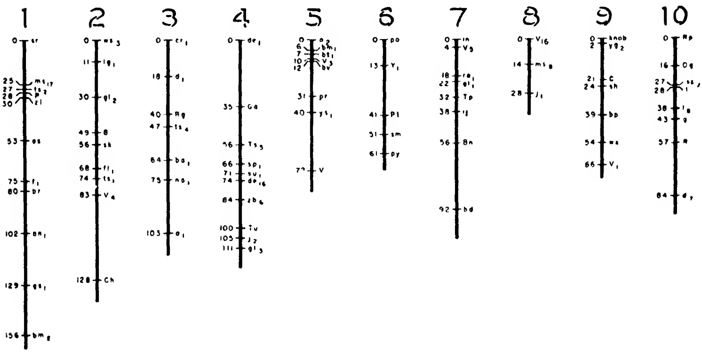


FIG. 4. The gene maps of the chromosomes of corn (maize), described on pages 8–10.

Complications arise with qualities which vary in their expression under different conditions, and with characters which depend on the interacting effects of various genes, but the inheritance of the single gene is the simplest process which can be imagined for biparental reproduction. It parallels the transmission of the chromosome segment concerned through the regular cycle of alternate doubling and halving of the chromosome complement in sexual reproduction.

Consider for example the genes *P* and *p* (Chromosome 1) which distinguish red-cob from white-cob corn. Each corn germ cell will have one or the other of these genes, not both, for each germ cell contains only a single set of chromosomes. Crossing a red and a white variety of corn is, in effect, simply adding to-

gether a germ cell of one and a germ cell of the other. The fertilized egg will contain both P and p. The corn plant which develops from this fertilized egg will have both P and p in each of its body cells, for these will be produced by repeated reduplication of the chromosomes of the fertilized egg. But the germ cells of this hybrid plant will have either P or p, and since the germ cells are produced by division of the Pp body cells, the two types of germ cells will be produced in equal numbers.

As far as P and p are concerned, therefore, there will be in the progeny of this hybrid corn if crossed with another hybrid of the same genetic constitution three kinds of plants: (*a*) PP plants, with red cobs and with only P germ cells; (*b*) pp plants, with white cobs and with only p germ cells; and (*c*) Pp plants, with red cobs and with both P and p germ cells. The relative proportions of these types of plants may easily be determined from the chance recombinations of the germ cells produced.

But other gene differences are likely also to be involved in the cross. Each of these will be inherited in the same simple fashion as P, but when many are involved together their combinations may be confusing. Genes carried by different chromosomes will be inherited independently, for the set of chromosomes received by any given germ cell will include maternal and paternal chromosomes in random combinations. For example, if the parents differed with respect to Lg as well as P, the hybrid would produce both Lg and lg germ cells, and the occurrence of Lg and lg would be independent of the occurrence of P and p. Accordingly there would be four types of germ cells produced in equal proportions by these plants: $P\ Lg$, $P\ lg$, $p\ Lg$, and $p\ lg$.

Genes carried by the same chromosome tend to remain in the parental combinations, but only to a limited extent. If the chromosomes were transmitted unchanged from generation to generation, two genes in the same chromosome, like P and sr in Chromosome 1, would always remain together—in fact, we would have no way of knowing that their effects were not due to a single gene. But regularly, in the production of germ cells, there is

an exchange of corresponding segments between corresponding paternal and maternal chromosomes. Consequently the chromosome received by the germ cell may be derived partly from one parent and partly from the other. Genes carried by the same chromosome therefore recombine, to an extent dependent upon their distance apart in the gene string. The frequency of this recombination is, in fact, the basis of the gene map.

With ten gene differences, over a thousand different gene combinations are possible in the germ cells; with twenty, over a million. The gene combinations possible in the progeny are all the possible combinations of the different germ cells of one parent with those of the other. Obviously the inheritance of any characteristic depending upon the interaction of several genes may be very complex. But the individual genes concerned, whenever they may be followed, show the same simple mode of distribution.

Our present view of the mechanism of heredity, then, is essentially a recognition of the existence within the cell of specific materials of characteristic function (the genes), and of a machinery for their orderly distribution through individual development and through successive generations (normal chromosome transmission). What these specific materials are and how their changes occur is almost wholly unknown. The distribution mechanism is more easily studied, and much progress has been made recently in our understanding of this phase of the problem.

Heredity may be altered by changing the germ cells, either by change in gene constitution or by change in chromosome organization.

THE GENETIC EFFECTS OF X RAYS

When chromosomes are treated with X rays, changes are produced which lead to the appearance of new hereditary characteristics in the progeny. A typical example is the occurrence of plants without chlorophyll, as illustrated in Figure 5. These plants are unable to carry on photosynthesis, and consequently

they die as soon as the food reserve in the seed is exhausted. Many other types of chlorophyll abnormality occur similarly as mutants. In some of these the affected seedlings are yellow, in others pale green, striped, or mottled. Many of the mutant variations previously known in corn have been produced anew by radiation treatment, and in addition there are many new types not previously found. The mutant types are not limited to chlorophyll variations; they include also morphological variants of

FIG. 5. Corn plants without chlorophyll (white), the result of a mutation induced by X rays.

all sorts—dwarfs, distortions, defective seed types, and so on. Since our study of the mutants has been carried on largely with plants not grown beyond the seedling stage, the types identified are largely those in which the seed and the seedling are affected.

When one of these types is established as a new race of corn, and is then genetically analyzed to locate the change in its germ plasm, the results are those to be expected from the substitution of a single mutant gene for the corresponding normal gene. In other words, the treatment appears to have changed a normal gene into one of its mutant forms.

But it soon became clear that the effects of X-ray treatment were not confined to gene changes. The cross shown in Figure 6 will serve to illustrate the point. The female parent has

the genes determining colorless endosperm (a_2), wrinkled surface (bt_1), and red aleurone (pr). The male parent has the corresponding genes for colored endosperm (A_2), smooth surface (Bt_1), and purple aleurone (Pr). The three genes may be found on the gene map (Figure 4) in Chromosome 5.

The hybrid ear without treatment would look like the male parent, since the genes *A*, *Bt*, and *Pr* are dominant. Each hybrid

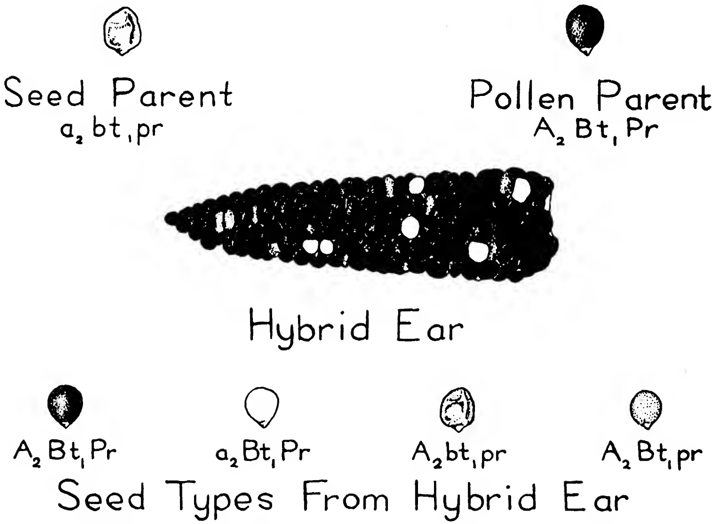


FIG. 6. Illustrating effects of X-ray treatment of corn pollen as described on pages 12–14.

seed is colored because it has a gene *A* from the male parent; if it lost this gene by mutation or otherwise it would not be colored. In the same way, losses of *Bt* and *Pr* may be identified. Thus the characters of the seeds serve as an index of the presence or absence of the corresponding genes in the male germ cell. Each seed is a test of the genetic changes which occurred in one treated pollen grain.

The hybrid ear produced by the use of X-rayed pollen is

shown in the figure. There are several colorless seeds, showing the loss of *A* in the pollen, several wrinkled seeds, showing the loss of *Bt*, and several red seeds, showing the loss of *Pr*.

But the significant point is that the wrinkled seeds are almost always red rather than purple. In other words, when *Bt* is lost *Pr* is almost always lost with it, though there are many cases in which *Pr* is lost without the loss of *Bt*. The loss of *A*, on the other hand, is not related to the loss of *Bt*.

In terms of the chromosome map, a simple explanation may be suggested. Assume that the gene *A* is on one arm of the chromosome, *Bt* and *Pr* on the other. The point of attachment of the chromosome in cell division is between *A* and *Bt*. When the chromosome is broken, that part which is severed from the attachment point is lost. Consequently a break between *A* and the attachment point results in the loss of *A* only; a break between *Bt* and this point results in the loss of *Bt* and *Pr* together; a break between *Bt* and *Pr* in the loss of *Pr* only.

If this is the correct explanation, many of the losses of genes in the treated germ cell must be due to actual loss of chromosome material rather than to changes in the genes. The red color of a *pr* seed in Figure 6 might be due to a change of *Pr* to *pr*, but alternatively it might be due simply to the loss of a segment of the chromosome including the *Pr* gene. In the *bt pr* seeds we may be sure that the red color is due to the loss of a segment including *Pr*, following a break between *A* and *Bt*. We therefore suspect that in some, at least, of the *pr* seeds the loss of the *Pr* gene is due to a break between *Bt* and *Pr*.

So far as endosperm changes are concerned, this explanation must remain purely hypothetical, for the genetic changes assumed cannot be checked either by progeny tests or by direct examination of the chromosomes. The endosperm has no progeny, and in its mature stages its chromosomes are most unfavorable for study. But analogous changes may be produced in the embryo, and may be identified by the changed characteristics of the resulting plants. Genetic tests may then be made with the

affected plant, and microscopic study may be made of its chromosomes. The cases are a little harder to find, since it is easier to grow and examine 10,000 endosperms (about thirty ears of corn) than to grow and examine 10,000 plants (about an acre of corn plants).

The type of analysis that may be applied may be illustrated by an example similar to the endosperm example just considered. Instead of genes affecting the endosperm, we shall use, as chromosome markers, genes affecting the characteristics of the plant, such as V_4 B Lg_1 in Chromosome 2. In the progeny from irradiated germ cells we find, as in the endosperm case, individuals in which one or more of the marker genes has disappeared.

Consider for example a liguleless (*lg*) plant, in which the gene *Lg* from the irradiated parent has been lost. Was the loss due to the transformation of the gene *Lg* to a new form in which it has the effect of the gene *lg*? Was it due to the loss of a piece of chromosome including the gene *Lg*? Or was it due to some other derangement of the chromosomes of such nature that the modified gene complex is no longer able to produce the *Lg* effect?

The constitution of the new liguleless plant may be investigated by breeding experiments, comparable to those on which the original gene map was based. The details of this analysis need not concern us here. Essentially it consists of the construction of a new gene map from the mutant individual. If the new gene map differs from the standard, the difference is the change induced by the X-ray treatment; if it is identical with the standard map, we can only assume, tentatively, that the treatment has transformed the *Lg* gene to the *lg* form. Such genetic analyses may be made of any of the abnormal plants occurring in the progeny from treated germ cells.

These genetic analyses of specific instances indicate various types of chromosomal derangement induced by the treatment. Sometimes the change is apparently the loss of a piece from the end of the chromosome. More often the piece lost is a segment not extending to the end, as if a section were cut out of the gene

string. Various other types of chromosomal derangement also are found among plants of the progeny produced from irradiated germ cells.

Some typical derangements found are illustrated diagrammatically in Figure 7.

A represents a simple deficiency including *Lg*. The normal chromosome is shown on the left. The modified chromosome following treatment still carried the genes V_4 and *B*, but it had lost *Lg* and also certain unidentified genes essential for normal pollen development. Presumably these unidentified genes were included with *Lg* in the segment lost.

C represents the loss of a nonterminal segment. The segment lost includes V_3, but the modified chromosome still carries the genes A_2 and *Pr*, which are located on opposite sides of the deficient segment. This shows that the segment lost cannot be terminal, while in A there is no genetic evidence to determine whether the lost segment was terminal or not.

B represents a different type of change. Nothing was lost, but the order of genes in the normal chromosome, A_2-*Bt*-*Pr*-V_2, has been changed to A_2-*Pr*-*Bt*-V_2, as if an internal segment were cut out, turned around, and reattached (inverted) in its original position.

D involves an exchange of segments between two different chromosomes. The normal chromosomes in the standard stock before treatment are shown on the left, one black and one white. The modified chromosomes, after treatment, are shown on the right. A piece from the end of the black chromosome has changed places with a piece from the end of the white chromosome. This type of change is the most common of the chromosome derangements found following X-ray treatment.

E represents a less common type of transfer of a segment from one chromosome to another. In this case a nonterminal segment of the white chromosome is cut out and inserted in an internal position in the black chromosome.

All of the five types of derangement illustrated are found

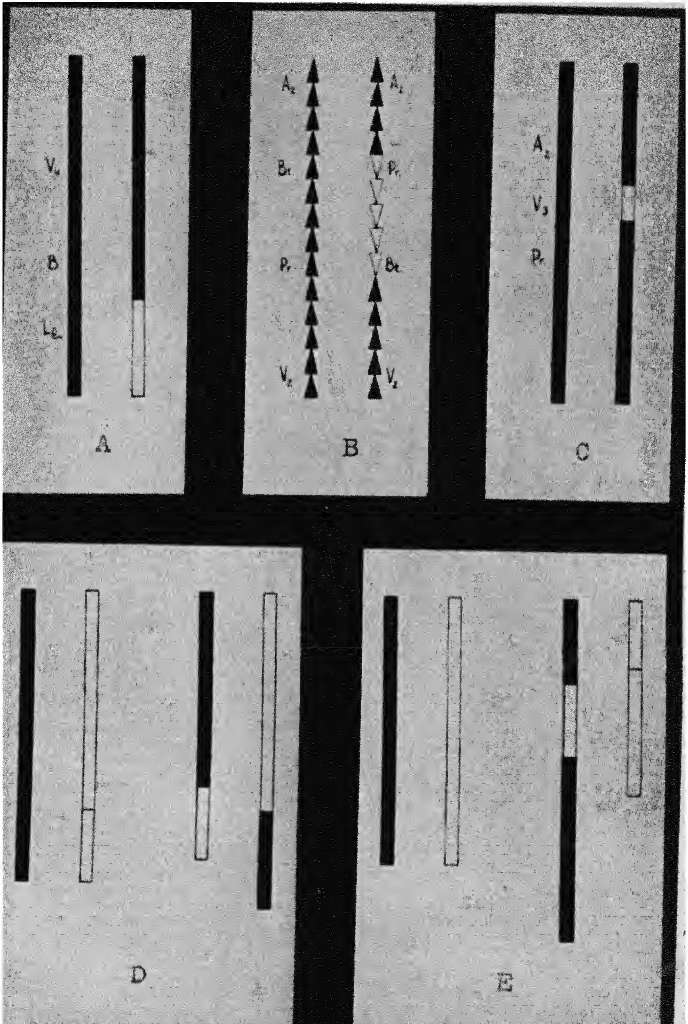


FIG. 7. Types of chromosome derangement in corn produced by X-ray treatment as determined by genetic analysis. A, terminal deficiency; B, inversion; C, nonterminal deficiency; D, interchange; E, intercalation.

repeatedly in progenies produced from X-rayed germ cells. In addition there are many instances in which a hereditary change is found with no detectable change in the gene map.

The chromosomal derangements diagrammed are, of course, hypothetical, for they are based wholly upon genetical analysis. A variant plant appears, and on the basis of breeding we construct a gene map which indicates that certain changes in the chromosomes have occurred. In other words, we predict that if the chromosomes could be examined in their fine structure, certain changes would be found. But few of these changes are large enough to be detectable in the small condensed chromosomes comparable to those illustrated in Figure 3.

To check the assumed derangement by direct observation of the changed chromosomes, it would be necessary to find conditions under which the chromosomes could be examined with far more precision. A most favorable stage for this purpose is the early preparatory stage of the cell division in which the chromosome complement is reduced from the double to the single set, for at this stage the corresponding gene strings must be fully extended and exactly paired to permit crossing over. This stage in corn was minutely studied by McClintock, who developed a technic for the identification of derangements involving the loss or transfer of the chromosome segments.

The appearance of the chromosomes at this stage is shown in Figure 8, which includes photographs of McClintock's preparations from plants with X-ray-induced alterations of various types. A represents an apparently terminal deficiency in the short arm of Chromosome 2. Genetic analysis of this plant showed that a segment including the gene *Lg* had been lost. The cytological preparation not only shows that a portion of chromosome is actually missing, but it indicates what part of the chromosome it is that has been lost. The gene *Lg* is thus shown to be located in this part of Chromosome 2. B shows the cytological configuration found in "inversions," such as the plant diagrammatically represented in Figure 7, B. If the chromosomes pair

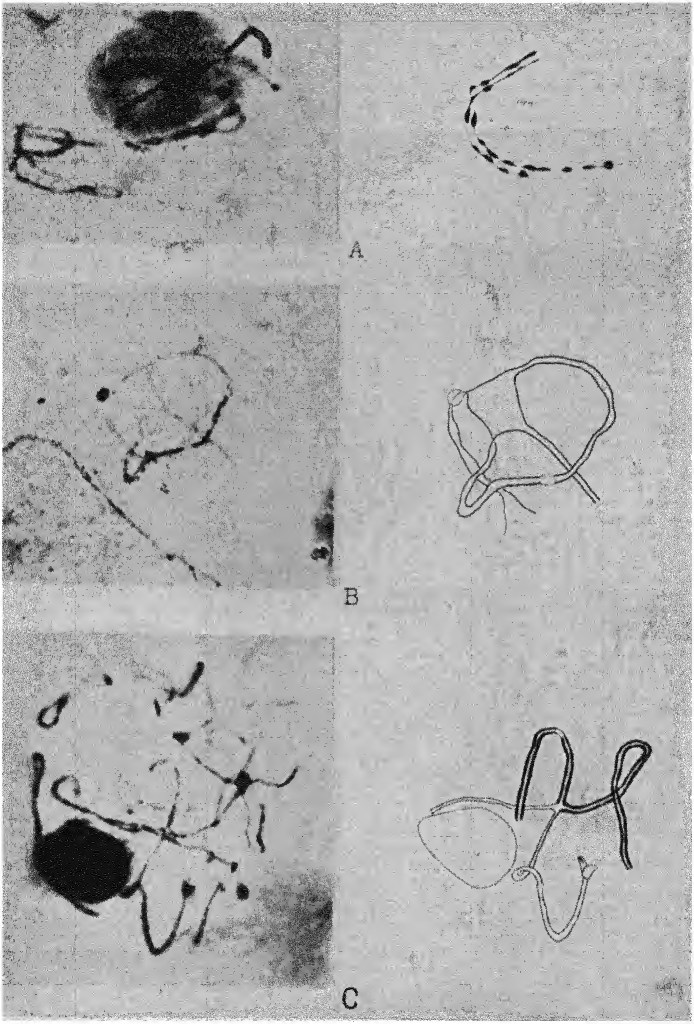


FIG. 8. Corn chromosomes showing various derangements induced by X ray. Photomicrographs, left; drawings, right. A, deficiency; B, inversion; C, interchange. (After McClintock.)

gene for gene, the only arrangement that will permit the pairing of corresponding regions throughout the length of the paired chromosomes is the formation of a loop at the position of the inverted segment, with the paired strands traversing the loop in opposite directions. C shows the appearance of the paired chromosome in an interchange plant, of the type diagrammatically represented in Figure 7, D. Here the pairing of all homologous

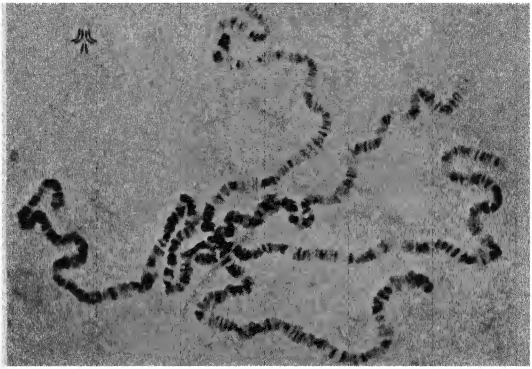


FIG. 9. Giant salivary gland chromosomes of Drosophila. Comparative size of the typical somatic chromosomes illustrated in Fig. 3. is shown in the upper left-hand corner. (After Kaufmann.)

regions can be accomplished only by the formation of a cross-figure, in which each of four chromosomes is paired with two partial homologues.

The development of this technic made it possible to investigate chromosomal derangements with far greater precision, to check the validity of the gene maps previously constructed on the basis of genetic evidence, and to place the genes in definite regions of the visible chromosomes.

A few years later a technic with the same advantages in far greater measure was developed for Drosophila chromosomes, chiefly through the work of Heitz and of Painter. The appearance

of the chromosomes of the salivary glands of the Drosophila larva is shown in Figure 9. The chromosomes of maternal and paternal origin are closely paired, so that deficiencies, inversions, and translocations may be readily identified. The chromosomes are enormously increased in size, and the characteristic transverse banding makes the identification of specific regions far more precise than is possible with the chromosomes of corn.

This permits the identification of the precise position of genes with reference to the visible bands. Figure 10 shows the position of a number of genes in Chromosome 1, in comparison with the genetic and the cytological gene map. The position of each gene

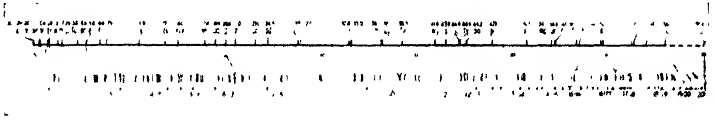


FIG. 10. The position of some of the genes of Chromosome 1 of Drosophila, as shown in the gene map (Fig. 3), located with reference to the visible bands of the giant salivary gland chromosome (below).

is determined by cytological study of mutant flies in which there has been a chromosomal derangement.

For example the white-eye gene is located in the region of the chromosome marked by a certain band in the section designated 3, C. Derangements involving the loss of the white-eye gene or the transfer of this gene to another position always show corresponding changes in this precise portion of the chromosome. The segment affected may include more or less material to the right or left of this region, and when it does the corresponding genes also are involved in the derangement. There is no implication that the band is the gene, but the position of the gene in the chromosome may be very narrowly localized to the region marked by this visible structure.

This minute study of the chromosome region in which a genetic change has occurred shows in some instances that what appeared to be a gene mutation was in reality a chromosomal derangement.

Sometimes a very short segment of the chromosome is lost, with effects which duplicate those to be expected from the mutation of a gene; sometimes an apparent mutation is due to the duplication of a region or to a changed order of genes in the gene string.

In a suspiciously large number of cases the apparent gene mutations turn out to be located at the exact point where a derangement has occurred. So many supposed mutations have been found to be due to extra-genic causes, that some geneticists suspect that orthodox gene mutation, in the sense of transformation of the gene, may not be involved in any of the alterations induced by X-ray treatment.

How do X rays produce alterations in heredity? As the typical cases diagrammed in Figure 7 show, profound changes in the chromosome may be produced by treatment of the germ cells with X rays. These changes include loss of segments of the gene string, inversion of segments, and translocation of segments from one chromosome to another. The losses are usually if not always nonterminal—in some cases cytological study indicates that the segment lost is terminal, but for technical reasons these cases are not conclusive. The inversions also are commonly if not always nonterminal. Translocations consist usually of the interchange of terminal segments, and less frequently of the intercalation of nonterminal segments. Apparently a translocated segment is never attached to a normal chromosome end; if terminal, it always takes a position where another segment has been removed; if nonterminal, it takes a nonterminal position.

The occurrence of these types, and absence of the missing types, suggests a possible mechanism. If it be assumed that the chromosomes when broken are capable of reattachment at the points of breakage, then the types observed are precisely the types which should occur and persist following chromosome breakage. Certain other types should occur but should be unable to persist through repeated cell division because of the absence of spindle attachment points or because of the presence of two spindle attachment points in the same chromosome. The plausible assump-

tion that the possession of a single spindle attachment point is a necessary requirement for normal chromosome behavior is confirmed by the study of actual cases. It is reasonable to assume therefore that the chromosomal derangements are the result of breakage followed by reattachment of the fragments at their broken ends. Reattachment of fragments in their original order results in chromosomes indistinguishable from the original type, or may perhaps yield chromosomes with mutations or short deficiencies at the points of breakage. Reattachment in new combinations may yield interchanges, intercalations, inversions, or internal deficiencies, depending upon the relations of the fragments reattached. Terminal deficiencies may result from breaks not followed by reattachment. Terminal deficiencies, however, are very rare in Drosophila, and the evidence for their occurrence in maize is not entirely convincing. There are many cases, like that illustrated in Figure 8, A, which appear to be terminal but which may be nonterminal deficiencies in which one break is very near the end of the chromosome.

An alternative possibility, that the derangements result from interchange of segments at points of contact between strands, is essentially a special case of the first hypothesis, in which it is assumed that the break must regularly and immediately be followed by reattachment in a new combination. This is not positively excluded by the present evidence, but it seems probable that reattachment is not necessarily immediate, and in fact that reattachment may sometimes fail to occur.

So much for the induced "chromosomal derangements." What of the induced "gene mutations"?

Among the variants induced by X-ray treatment there are many apparent gene mutations which are free from any chromosomal derangement detectable with our present methods. They appear to be identical in many instances with apparent gene mutations of spontaneous origin. In several cases, in Drosophila, it may be shown that a mutation induced by X rays is capable of reverse mutation to the original form, and that the reverse mu-

tation also may be induced by X-ray treatment. This virtually excludes the possibility that the induced mutation is due merely to the physical loss of the gene.

Nevertheless, so many cases have been found in which demonstrable extra-genic changes simulate gene mutation, that one is skeptical of the positive identification of any of the induced mutations as the result of a transformation in the gene itself. The demonstrated fact that changes in gene arrangement may simulate gene mutation is especially significant in this connection. For if the effect of a gene may be modified by a new neighboring gene brought into the neighborhood by a gross chromosomal derangement, it is reasonable to suppose that analogous effects may be produced when the new neighbor is brought in by a derangement on a smaller scale, perhaps below the limit of cytological detectability. Even the mutations which are capable of reversion may be of this class, for the reverse mutation may be due merely to a second derangement removing the neighbor responsible for the original disturbance.

THE GENETIC EFFECTS OF ULTRAVIOLET RADIATION

It is therefore possible that all of the genetic alterations induced by X rays are related in origin. The deficiencies, interchanges, inversions, and intercalations may be merely the result of varying reattachments following chromosome breakage; the mutations may be merely the phenotypic results of gene losses, duplications, and rearrangements incidental to the chromosomal derangements.

On the contrary, it is possible that the diverse genetic alterations may include types quite distinct in origin, representing changes fundamentally different in nature. Their association in X-ray progenies may be due only to the fact that the radiation is a powerful agent which may be capable of affecting all of the diverse reactions involved.

If the former is true, we may predict that somewhere in the

spectrum, between the genetically effective X-ray region and the genetically ineffective visible light region, a point may be found at which genetic effectiveness ceases, and at this point all of these effects will disappear together. If on the contrary the associated effects are due to independent reactions affected by X rays, it is possible that their spectral relations may be quite different. In this case, points in the spectrum may be found beyond which some of the effects are produced while others are not. This would permit a more or less selective alteration of the genotype and might make possible the production of mutations without the accompanying chromosomal derangements. It is even possible that spectral relations might be so specific as to permit the selective transmutation of particular genes.

All of the X-ray-induced alterations which have been mentioned may be classified under three heads: mutations, deficiencies, and translocations. There is in addition a type of alteration regularly found which may not belong to any one of these three classes. This is the so-called "dominant lethal"—that is, the abortion of the embryo produced by the fertilization of eggs with irradiated male germ cells. Since the individual affected is dead before it can be studied, no genetic or cytological analysis is possible, and we cannot determine the cause of the abortion. It may be assumed that these individuals represent germ cells in which a dominant lethal gene was produced by the treatment, or that they are the result of the loss of chromosome regions so essential that the embryo is unable to develop without them. They would then belong to the class of mutations or of deficiencies. But since they are beyond the reach of analysis the possibility must be considered that they are the result of other induced changes in the male germ cell which prevent the development of the embryo.

All four types of genetic change mentioned—mutation, deficiency, translocation, and embryo abortion—are found in all X-ray experiments reported in which the conditions were such as to permit their detection. This is true regardless of the X-ray

wave lengths used, although most of the evidence for the longer X-ray wave lengths comes from experiments in which the mutational effect alone was under study.

Our first question regarding the effects of the ultraviolet, then, is whether this association of the four types of genetic alteration is maintained up to the point where all genetic effectiveness ceases.

Another reason for studying the genetic effects of different wave lengths of ultraviolet radiation is the possibility of finding some clue to the chemical nature of the substance absorbing the radiation that produces the various genetic alterations. Ultraviolet absorption, unlike X-ray absorption, is dependent upon molecular constitution. Many chemical compounds have distinctive absorption spectra in the ultraviolet, that is, they absorb certain wave lengths much more readily than others.

This makes possible a method of study which we may call the shadowing method. Your shadow is the patch of darkness which you produce by intercepting and absorbing the light. If you were perfectly transparent to light waves you would have no shadow. Actually you are relatively transparent to some varieties of "invisible light." If you swallow a safety pin, the doctor will make use of this fact. He will photograph you, using the invisible light of an X-ray lamp, or rather he will photograph the shadow which you cast in this light. Your flesh does not absorb much of this radiation, your bones absorb more, and the metal pin absorbs still more. The result is a photograph showing the shadow of the pin in place among the shadows of your bones and internal organs (Fig. 11).

If you pass white light through a layer of water, the emerging light will still be white light. If now you dissolve a little copper sulphate in the water, the light emerging will be blue. The white light is a mixture of wave lengths, of all the colors of the rainbow. The water absorbed these various wave lengths almost indiscriminately, and therefore made little change in their proportions in the mixture. But the copper sulphate solu-

tion absorbed them selectively, absorbing relatively much of the red, for example, and relatively little of the blue. It would cast a dense shadow in red light, but a faint shadow in blue light. In white light it casts a blue shadow. Many other substances behave in the same way, absorbing light of different colors. Thus we have a kind of qualitative shadowing, depending on the chemical constitution of the absorbing substance. It

FIG. 11. Shadowing a mislaid safety pin in the chest region with the help of X rays as described on page 26.

is possible to determine quite precisely the wave lengths concerned and the extent to which each is absorbed per molecule of the absorbing material. This may be done not only with the visible light but also with the invisible ultraviolet and infrared radiation. The absorption spectra thus determined turn out to be quite characteristic for many substances.

Now the radiation which produces the genetic alterations must be radiation which is absorbed by the chromosomes, and the relative effectiveness of the ultraviolet wave lengths may indicate the relative absorption of these wave lengths by the substance responsible. Correlation of this evidence with the absorption spectra of substances known to be present in the chromosome, or of substances suspected of playing some part

in genic reactions, may yield some circumstantial evidence regarding the constitution of the germinal substance.

The chemical composition of the chromosome is not fully known, but we do know that it is made up in large part of nucleic acid combined with a protein or proteins. Both the nucleic acid and the protein constituent have characteristic absorption spectra in the ultraviolet. The absorption spectrum of nucleic acid, as determined by Caspersson, is shown in Figure 12.

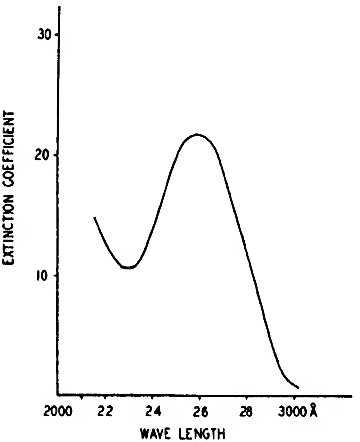


FIG. 12. The relative absorption of different ultraviolet wave lengths by nucleic acid. (After Caspersson.)

If the radiation which produced chromosome breaks is absorbed by nucleic acid, then the effectiveness of different wave lengths in producing chromosome breaks should vary as indicated by this curve. Similarly, if the effective radiation is absorbed by the protein constituent, the relative effectiveness of the wave lengths should vary according to a different but equally distinctive pattern. If mutations are due to absorption of radiation by different substances, their spectral response may be distinguishably different, and may perhaps yield some suggestion of the substance involved. We may thus apply the shadowing method, so to speak, to the substances involved in the genetic alterations.

When the pollen is treated with ultraviolet radiation, the effects are superficially similar to those following treatment with X rays. Many deficiencies are found in the resulting seeds and in the plants which are grown from them. Embryo abortion occurs in a considerable proportion of these seeds. The mutation rate is greatly increased.

A striking difference is found when genetic analysis is applied to the plants produced by the use of the irradiated pollen. Although deficiencies are found among these plants, as with X-ray treatment, translocations are very rare.

It is clear therefore that the "X-ray complex" of genetic alterations may be broken down. Deficiencies and mutations may occur without the accompanying translocations. This does not necessarily mean that translocation is unrelated to deficiency, for there may be types of deficiency which also are absent or rare in the ultraviolet progenies. But there must be types of deficiency which are not dependent for their occurrence upon a translocation mechanism.

This result shows also that the induced mutations are not merely the result of changes in the position of genes, as has been inferred by some geneticists on the basis of the X-ray results. Here is a treatment which produces mutations as frequently as does X-ray treatment, and which has little or no effect upon the occurrence of translocation.

When the shorter wave lengths of the ultraviolet range are filtered out, another difference appears. Embryo abortion becomes quite infrequent without a corresponding drop in the frequency of deficiency and mutation. If the aborted embryos are due wholly to deficiencies and mutations, there should be a constant relation between their frequency and that of deficiency and mutation, regardless of wave length. This is decidedly not the case. Some of the aborted embryos may be due to deficiency—in fact, from a consideration of the frequency and the effects of the deficiencies observed it seems certain that some combinations of deficiencies must occur which must be lethal to the embryo. But the comparisons of the effects of different wave lengths indicate that by far the greater part of the aborted embryos are due to some other effect of the radiation, an effect much more pronounced at the shorter wave lengths than at the longer.

Wave lengths longer than 290 mμ have relatively little

effect on the frequency of embryo abortion, but wave lengths up to beyond 300 mμ show definite effects on the occurrence of deficiency and mutation. There are two strong lines in the mercury spectrum within this range, λ297 and λ302. Their effect on deficiency and mutation is less per unit of dose than that of some shorter wave lengths, but these radiations are tolerated in much heavier doses. Using heavy dosage it is possible to obtain with these wave lengths the highest frequencies of deficiencies and mutations that can be produced with any of the ultraviolet radiations available. The next wave length, λ313 has no appreciable effect, although even higher doses of this may be applied to the pollen. Thus deficiency and mutation appear to have the same or very similar wave-length relations. Further study is needed to determine whether any difference in spectral response may be shown for these effects in general, or for specific mutations or deficiencies.

Wave lengths longer than λ313 apparently do not affect any of the four types of genetic alterations mentioned. By the use of filtered ultraviolet radiation it is possible to apply very heavy doses of mixed wave lengths extending from λ313 into the visible range. This does not produce any significant increase in the frequency of genetic alterations. Thus the wave-length limit for the genetic effects of radiation appears to be between λ302 and λ313. There is of course for the longer wave lengths the possibility of very faint effectiveness which could not be detected except in special experiments.

The relative effectiveness of the different wave lengths in producing each of the genetic alterations may be determined by applying pure (monochromatic) radiation of each wave length in equal doses to separate specimens of pollen. By the use of special equipment it is possible to apply these treatments to quantities of pollen sufficient for effective pollination, and in the case of the endosperm deficiencies at least the genetic effects are frequent enough to permit making significant comparisons.

But such comparisons do not represent the effectiveness of

the different wave lengths in terms suitable for comparison with specific absorption spectra. The equal doses compared are doses equal in terms of the energy which reaches the outside of the pollen grain; the energy which reaches the chromosomes, or the points within the chromosomes where the energy is absorbed, may be quite unequal.

The pollen grain is roughly spherical in shape, and about 95μ in diameter. There are two sperms, one of which will fuse with the egg to form the embryo whereas the other will fuse with two nuclei of maternal origin to form the endosperm. The two sperms are indistinguishable in appearance, and are eccentrically located in the mature pollen grain. The pollen grains lie in a single layer for treatment, but if they are oriented at random it is obvious that the sperms will be much nearer the upper surface in some pollen grains than in others. This is a factor of little importance with X-ray treatment, since the radiation is so little absorbed that the intensity at different points in the pollen grain is not very different. But with ultraviolet radiation, due to its high absorption, the dose applied to the sperm nucleus in a favorably oriented pollen grain is very much higher than that applied in an unfavorably oriented grain of the same treated sample. The result is a high frequency of coincidence of independent effects (since the grains most likely, because of their orientation, to be affected by one change are the ones most likely to be affected by another) and a much flattened dosage curve for specific effects (since the first increments of dose tend to affect the most favorably oriented pollen grains, and

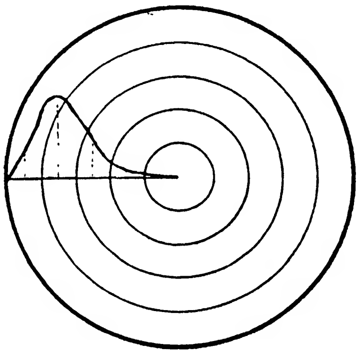


FIG. 13. The location of the sperms in the corn pollen grain, shown as frequency distribution in five concentric shells. (From Cameron.)

successively later increments must produce their effects in less and less favorably oriented ones). The maximum frequency of specific deficiencies attainable by increasing dosage may be almost doubled (at λ254) by applying the radiation half from above and half from below the pollen. This internal filtration introduces serious complications in the interpretation of the wave-length comparisons, which cannot be fully discussed here. (Fig. 13.)

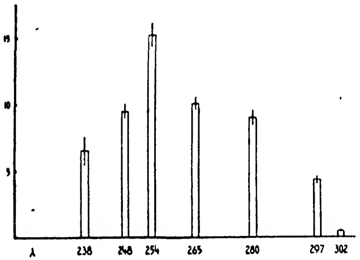


FIG. 14. The relative frequency of endosperm deficiencies induced by pollen treatment with various ultraviolet wave lengths. The height of the columns indicates the percentage of deficiencies for the genes *A*, *Pr*, and *Su*; the length of the vertical line at the top of each column, the probable error of determination.

A comparison of the effectiveness of various ultraviolet wave lengths in inducing endosperm deficiencies is shown in Figure 14. The doses applied were equal in energy incident at the surface of the pollen grain (2,000 ergs per square millimeter). The curve shows a very suggestive resemblance to the absorption spectrum of nucleic acid (Fig. 12). In view of the complications from internal filtration which have been mentioned, the significance of this resemblance can be determined only by further study.

II

THE REGULATION OF PLANT GROWTH

By F. W. Went

California Institute of Technology

The first three decades of our century have shown a progressive specialization of the biological sciences into a large number of separate disciplines, each with its own methods and aims. This has resulted in a very rapid development of each branch, but biology as a whole has suffered. Investigators could no longer see an organism as a unit; the physiology of some phenomenon or reaction was studied without proper knowledge of the structure of the system involved; plant diseases were studied without sufficient knowledge of the normal plant; geneticists felt more akin to mathematicians than to fellow biologists; and taxonomists disregarded all other biological disciplines. Specialization seemed also to stress the differences rather than the similarities between plants and animals.

All this is changing, or has changed, and no better example of this changed attitude can be given than the recent development of the field of growth hormones of plants. Not only is there close coöperation between a large number of laboratories here and in Europe, but also within one institute a number of investigators, chemists, plant physiologists, geneticists, plant pathologists, and plant anatomists have come to work out problems together. I hope that this discussion will convey the impression that our attack on the how and why of growth is such a coöperation on a very wide basis. My aim is to present a synthetic picture of a plant in its growth and development as far as our present knowledge allows.

It will be impossible, of course, to present all the experimental detail on which this picture is based, and many important angles of the problem of growth will not be mentioned. [1] *

Growth, in general, is only possible within a rather narrow range of suitable external conditions such as the right temperature, a medium containing sufficient water, and certain electrolytes in proper proportions. Although the plant is able to some extent to create its own surroundings for its growing parts, still any change in external conditions will be reflected in growth. But in the present discussion we shall consider only the internal regulation of plant growth, and therefore can disregard the influences of the external conditions by considering growth only under constant suitable external conditions.

Before it is possible to arrive at a synthesis we first have to analyze the growth process into its component parts. The first and most important step in the analysis of a plant was the cell theory. It gave us the smallest unit which still possessed the properties we associate with life and growth. For many functions and processes the cell was recognized as the smallest structural and functional unit, but after the first fifty years of the cell theory it became recognized that an organism is *more* than a colony of cells. The cells are all interrelated and interconnected in such a way that a collection of cells becomes an organism. It was found that different parts of the plant were dependent upon each other, which was expressed by the new concept of correlation. This first approach to a synthetic view of a plant reached its full development with the hormone concept of plant growth. The hormone concept transcends the differentiation of the organism into cells; hormones integrate cells into an organism.

There are two principles along which a coöperation between the parts of an organism can be achieved. One is a superregulating center having remarkable integrating properties and directing the activities of all parts of the organism. Our brain is such a

* Sources will be found in the list of References, p. 287.

center which can best be compared with the functioning of a totalitarian state. However, in a plant no nerves have ever been found and the integration of parts must be achieved by some other mechanism. As we shall see, the development of a plant is based on the democratic principle. As long as cells are self-sufficient they remain independent, and even when the cells remain attached no unity results. The organism starts only when one cell or cell group becomes dependent upon other cells for its development or functions. This differentiation is closely connected with the evolution of an organism.

I shall repeatedly use the term *hormone*. A hormone is a substance, or chemical messenger, which, transported through the body, determines the growth or function of parts remote from its place of origin. Usually it acts in minute quantities. At the same time this property of effect at a distance makes hormones easily accessible to experimental attack. We have only to intercept them during their transport.

Let us start by considering a very recent development in plant hormones, root growth, since it provides the simplest example of plant hormones. It is evident that a root cannot grow just by itself; it has to obtain its food from the assimilating parts of the plant above ground. Thus ordinarily a food correlation between roots and leaves exists. We can now ask, how independent are these different parts, or in more concrete terms, are roots able to grow by themselves in a medium containing water, salts, and sugar? In early experiments it was found that this was not the case. But Robbins [2] showed that with the addition of yeast extract to the medium much better growth of isolated roots was obtained. And it was left for White [3] to show that yeast extract contains all that is necessary to obtain unlimited growth of isolated root tips in a medium consisting of sugar and the necessary salts. Then simultaneously Bonner [4] and Robbins [5] discovered that vitamin B_1 was the most important ingredient of the yeast extract. Thus it is possible to grow flax roots indefinitely

in a medium containing salts, sugar, and vitamin B_1. However, with other roots it was found necessary to add still other compounds in minute amounts before unlimited growth could be obtained. Thus pea roots require nicotinic acid in addition to vitamin B_1 [6] and tomato roots thrive without nicotinic acid but with vitamin B_6. [7]

According to the evidence just presented the root system has a remarkable synthetic ability and is able to synthesize proteins, enzymes, nucleic acids, etc., using sugar and nitrate as starting materials. This means that flax roots for instance can synthesize purine, sterole, imidazole, indole, flavin, and scores of other organic nuclei. On the other hand they have completely lost the ability to synthesize pyrimidine and thiazole nuclei. This ability, however, is possessed by the above-ground parts of the flax plant and, therefore, the root system of a flax plant grows at the discretion of the top. Thus the utilization of sugars by the flax root is completely regulated by the supply of vitamin B_1. Other roots have lost the ability to form still other substances and therefore have become still more dependent upon the above-ground parts, but there is no difference in principle. We can imagine any number of combinations of substances which cannot be synthesized by root cells, and future work will show what variations on this general theme nature has realized. But as far as information is available we can say that vitamin B_1 is a general root-growth hormone of the higher plants and is produced in the leaves in light.

It is interesting to point out that all these growth factors, which evidently are hormones for the plants, are vitamins for animals. This means that the animal organism has completely lost the ability to form these substances and therefore has to derive them from its food. In plants this process has not progressed so far, and certain cells are still able to produce them. This shows how closely vitamins and hormones are connected and it also indicates the essential unity between plants and animals. It now becomes evident why plants are our best sources of vitamins. They do not

produce them for the sake of charity, but they are essential factors for their own lives.

Since we know the functions which, for example, vitamin B_1 or nicotinic acid perform in the cell (they are co-enzymes both in plants and animals), it will not be necessary to dwell on them any longer. It is clear that if no enzymes are present no enzymatic activity in the cell can occur, and the cell cannot function. These substances do not tell us more about the growth process itself; they only give us a better understanding of correlations in the plant and how these are brought about. Therefore we shall take up now another growth factor the metabolic function of which is not known as yet. This is auxin, the hormone for stem growth.

It was known for a long time that the stem tip is essential for the elongation of the stem below the tip. Since this subapical stem region is the part where the most rapid growth occurs in a plant, this correlation between stem tip and elongation is very pronounced and easy to investigate. As an example the oat seedling may be considered. When the oat seed germinates it first sends down some roots. This is closely followed by the development of a shoot consisting mainly of a hollow cylinder, the coleoptile, which encloses the first leaf. The coleoptile grows very straight, and three days after germination at 25° C and 85 per cent humidity it will have reached a length of approximately one inch (Fig. 15, A). When the tip of such a coleoptile is cut off the growth of its lower region diminishes and ultimately stops (B). This is a correlation phenomenon and not due to wound shock, since after replacement of this tip growth is resumed (C). In connection with the nature of this correlation effect between tip and growing region the following experiments are of importance.

Both Boysen-Jensen [8] and Paal [9] demonstrated that the effect of the coleoptile tip could be transmitted not only across the wound gap, but also across a thin layer of gelatin. This made it clear that the effect of the tip is due to some diffusible agent, whose existence actually was demonstrated by the following

experiments. Coleoptile tips were cut off and placed together on a thin layer of gelatin or agar-agar (F). After one or two hours the tips were removed, and the jelly on which they had stood was tested for its growth activity. If small blocks of such agar were placed on the stumps of the decapitated plants their growth was immediately accelerated (D), whereas no growth increase resulted from the application of pure agar blocks. A quantitative method for determining the amounts of growth-promoting material in the agar was worked out by applying small blocks only to one half of the cut surface of decapitated coleoptiles. [10] Since there is no lateral spreading of the growth-promoting material, only the side below the agar block will show increased growth which results in a curvature (E). It was found that within limits this curvature is directly proportional to the concentration of applied growth substance. When each determination of growth-promoting activity is carried out on at least 12 plants, significant data can be collected. This method of assaying growth-promoting substances is called the Avena (oat) test.

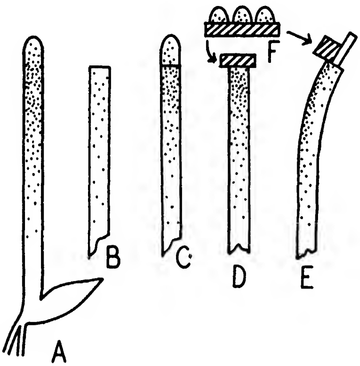


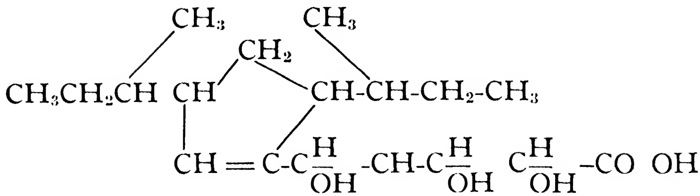
FIG. 15. Diagrams illustrating the effect of auxin, formed in the coleoptile tip, on the growth of the Avena (oat) coleoptile. Growth rate in various regions indicated by density of stippling. A, normal oat seedling; B, decapitated plant, growth rate greatly decreased; C, decapitated plant, but tip replaced, growth rate only slightly below normal; D, decapitated plant with agar block, on which tips stood for two hours, growth rate like C; E, decapitated plant with unilaterally applied agar block containing auxin, only the tissue below the agar block has an increased growth rate; F, diffusion of auxin from coleoptile tips into agar.

With this method a number of problems can be investigated, such as the place, rate, and conditions of growth-substance production, its behavior in the plant, and its chemical and physical

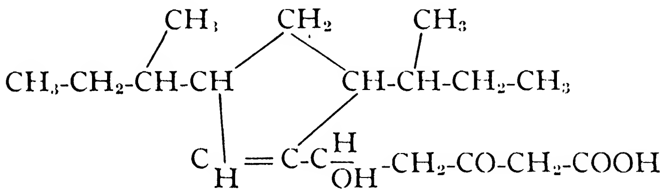
properties. Concerning the latter it was found that the growth substance coming from the cells of the coleoptile tips was heat and light stable, and its diffusion rate indicated a molecular weight between 300 and 400. [10] This was later verified after Kögl and Haagen-Smit had succeeded in purifying the growth substance and obtained it in crystalline form. [11] This purification of growth substances is one of the most brilliant chapters of organic chemistry, but it can be mentioned only very briefly here. Kögl and Haagen-Smit set out to collect the growth substance diffusing from coleoptile tips, but they gave up this approach when they calculated that it would take ten girls working ten hours per day for a period of seventy years to collect one gram of this material. Soon they discovered that human urine was an excellent source of growth substance, and after working up hundreds of liters of urine they obtained a couple of hundred milligrams of crystalline material which they called auxin *a* and which has the empirical formula $C_{18}H_{32}O_5$. Although it appeared that urine contained auxin as a result of plant food ingested, Kögl and Haagen-Smit were not satisfied until they had isolated auxin from a plant source as well. They succeeded in crystallizing auxin from germinating barley and from corn-germ oil, in both cases obtaining some auxin *a* in addition to auxin *b* ($C_{18}H_{30}O_4$). After a splendid chemical analysis they established the two structural formulas for auxin *a* and auxin *b* (page 40).

It is interesting to note that in the animal female sex hormones the two naturally occurring products (theelin and theelol) also differ by the elements of water—a difference due to the presence of two hydroxyl groups in the one against one keto group in the other.

The very small amounts of auxin obtainable by chemical extraction, as well as its auto-inactivation (even in crystalline form) made this auxin very difficult to obtain and to handle for physiological investigations. But here another discovery of Kögl and Haagen-Smit brought the growth substances into the realm of practical experimentation. They found that a completely different



Auxin *a*



Auxin *b*

substance, indoleacetic acid, has almost the same physiological activity in promoting growth in stems. [12] Although this was very nice from a practical standpoint, since indoleacetic acid is easily synthesized, it brought up a serious theoretical difficulty. At first we had considered auxin *a* and auxin *b* to be the specific substances inducing growth by cell elongation and now suddenly it seemed that this agent was not so specific after all. To escape this difficulty Kögl suggested the following adaptation of Emil Fischer's lock-and-key simile. If we consider that auxin *a* opens the lock of the door to growth then indoleacetic acid might either be a passkey for that same lock, or a key to a completely different lock, say for a back door to growth. Since Kögl did not like the idea of passkeys which would seriously affect our concept of chemical specificity in physiological processes, he preferred the idea of the existence of a back door to growth, a sort of double insurance such as exists in many physiological processes in animals. The subsequent discovery of so many more-or-less related substances (such as cis-cinnamic acid, naphthaleneacetic acid, phenylacetic acid, etc.) able to partake in the growth reaction did not

support Kögl's choice, but neither was it in favor of any specificity of the growth reaction. This leads to the theory that the action of all these substances was of the nature of a stimulation. Growth was considered to be ready to proceed as soon as it was set off by the trigger action of one of these substances. This view was strengthened by the fact that the effective concentrations of the different substances varied widely; one substance required a ten thousandfold higher concentration than another to produce the same growth effect. Thus the specificity of the growth reaction was reduced to some very unspecific stimulation. This unsatisfactory state of affairs will be analyzed in detail, for it is typical of many biological problems, and at the same time it touches certain fundamental principles of biology.

But first we have to consider some of the functions of auxins [1] in the plant. Their detection and analysis are based entirely on the effect they have on growth by cell elongation. They increase the plasticity of the cell wall, which then will yield to the turgor pressure. In this way auxin controls the growth of stems, leaf stalks, and flower stalks, though not that of roots. Since the responses of plants to gravity and light, which we call geotropism and phototropism, are due to unequal growth responses, we can inquire to what extent auxin controls these growth movements.

Dolk proved that gravity had no effect whatsoever on the production of auxin: [13] just as much auxin was produced in horizontal as in vertically placed stem tips. He found, however, that in horizontal stems the auxin which normally moved equally along all sides of the stem was deflected so that much more reached the under side than the upper. The lower side responded with an increased growth causing the stem to bend upward until the tip was vertical, and the auxin distribution had become equal on all sides again. In this case gravity caused deflection of the

1. Although it is likely that in the higher plants either auxin *a* or auxin *b* performs the functions of a growth hormone, not enough evidence is available to prove this beyond doubt. Therefore the name auxin will be used as a generic term, including all substances which have a definite growth-promoting effect on stem growth as measured in the Avena test (p. 38).

transported auxin. Similarly there was a measurable deflection of the auxin stream toward the dark side of a unilaterally illuminated stem; and the light destroyed more auxin on the illuminated than on the dark side. Both of these phenomena tended to decrease the auxin content at the illuminated side when compared with the dark side and together they explained the curvatures of stems toward the light. These phototropic and geotropic curvatures caused by auxin are essentially effects on growth in length, so that in both these cases auxin performs the same function. By the effect of gravity and light on their distribution the auxins regulate the position of the plants in space in response to their environment.

But rather soon after their discovery the auxins were found to perform other functions besides regulating growth in length. In investigations of the factors involving root formation on stems it was soon found that here also correlation carriers of the hormone type were involved. [14] This was indicated, for example, by the effects of leaves and buds on the root formation of stem cuttings. When a number of segments of a stem carrying leaves and buds are stuck upright in wet sand, roots will appear in a couple of weeks. But when the leaves and buds are removed no root formation occurs. Sugars, it was shown, were not responsible for the rooting response, but minute amounts of materials formed in leaves and buds moving down inside the stem and accumulating at the basal cut surface. In many plants this material was found to be identical with auxin. [15] Therefore root formation can be induced by an auxin treatment in cuttings without leaves and buds which normally will not root. In other plants it was found that not auxin but certain other factors were lacking to produce satisfactory root formation. In some seedlings this is sugar. In others root formation occurs but the root primordia formed do not grow out, due to lack of vitamin B_1. In such cases vitamin B_1 application is required for visible root formation on cuttings. [16] All these facts do not mean that root formation is unspecific since so many different substances can induce it, but

only that it is the result of many different processes and reactions, each one of which may be limiting and may be affected by other substances. This also means that, in general, a thorough investigation is required before conclusions concerning the lack of physiological specificity can be drawn.

Only one more auxin-controlled phenomenon will be mentioned, namely, the development of the ovary of a flower into a fruit. Ordinarily this only occurs after pollination and fertilization of the egg cell in the ovules. Failure of fertilization generally results in abscission of the flower. But Gustafson found that normal fruit would develop without pollination if only a small amount of auxin were applied to the pistil. [17] Such fruits, of course, lack seeds but they may develop to normal size. In this way so-called parthenocarpic tomatoes, bell peppers, cucumbers, and even watermelons have been produced. Apparently the applied auxin replaces the auxin normally supplied by the developing embryo. It seems that a deficiency in auxin production during the development of the fruit may lead to premature dropping, for Gardner has succeeded in preventing premature dropping of apples by spraying the trees just prior to the drop with an auxin. [18]

From the previous discussion it has become clear that many different substances may affect a number of different physiological processes. The effective concentration for each of these processes is different for most of the active substances so that every fact enumerated above pointed toward a stimulative action of auxin, and it seemed impossible to maintain the original thesis that auxin was the specific agent inducing the growth reaction. But at this point Thimann made an interesting discovery. [19] He showed that the activity of various substances differs in different tests. In all cases the activity in the pea test showed less quantitative variation than in the Avena test. The pea test is a method of determining growth substance activity by immersing split pea stems in the solution to be tested so that the substances can enter the plant tissues from all sides. The Avena test, which has been

described earlier, and which consists of the unilateral application of agar blocks containing growth promoting material to the cut surface of decapitated oat coleoptiles, requires more than simple penetration of the substances into the cells. As Thimann pointed out, only those substances which are capable of translocation downward in the coleoptile will reach the growing cells and only after successful completion of this transport will they be able to induce growth and curvatures. This means that a substance must possess at least two sets of properties, namely, those required for transport and those for growth, before it will be able to induce the Avena test or, put in other words, before it shows physiological activity. Therefore we must differentiate between all the different properties of a substance required for or influencing activity and then study *only* the specific property about which we want to get more data, eliminating the secondary properties, before any conclusions can be drawn concerning the activity per molecule of the substance in a definite physiological reaction.

In the Avena test the range of activities of a set of seven given substances was about ten thousandfold (Fig. 16). If these same substances were used in the pea test, which is also a test for growth activity, but which does not require longitudinal transport of the substances inside the tissues, then the range of activities was reduced to a hundredfold. The most remarkable fact in this connection was that the molar activity of a number of substances became equal but that still a certain number of others had lower activities. It was found that the activity of these growth-promoting substances in the pea test is due to two successive reactions in which they take a part. The first, or preparatory reaction, only conditions the cells for the actual growth reaction which follows, but does not by itself lead to measurable growth. It does, however, limit the growth activity of most substances. Therefore, if we test the various substances on peas in which the preparatory reaction is practically completed, other molar activities for the growth reaction proper are obtained which closer approach

equality. Since it was further known that only the undissociated auxin molecules were active in producing growth, a further correction for the different acid strength of the substances was made, comparing only the activity of the number of undissociated mole-

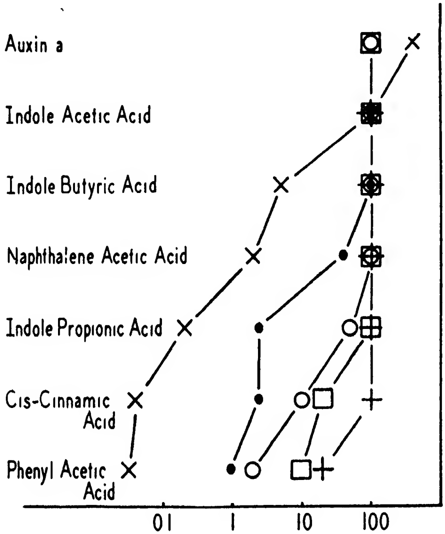


FIG. 16. Illustrating the relative activities (abscissa) of various growth-promoting substances. Activity of indoleacetic acid is always taken as 100.

× Activity in standard Avena test.
● " when tested on Avena coleoptile sections.
○ " in standard pea test.
□ " in pea test when peas have been pretreated for two hours with γ-phenylbutyric acid 100 mg./l. (activity in growth reaction proper).
\+ " in pretreated peas, but corrected for undissociated free acid molecules.

cules at the pH of the cell. This brought the activity of cis-cinnamic acid up to unity and left only phenylacetic acid below that of the other substances. This means that in the actual growth reaction all active substances take part in the same molar con-

centrations and that, per active molecule, the same amount of growth occurs independently of the chemical composition of the active substances. [20]

A difficulty which has to be faced before acceptance of this conclusion is that substances of such different chemical structure must all be able to take part in the same reaction. Therefore it was attempted to determine the common denominator for all these substances and such a common denominator was actually found. [21] All active substances contain an unsaturated ring system with a side chain of at least two carbon atoms adjacent to the double bond and a carboxyl group or a group readily hydrolysable into a carboxyl group in the side chain. Finally a certain special configuration between carboxyl group and ring is required. This shows that there is a very definite structural specificity of the substances which are able to partake in the growth reaction. This chemical specificity is not dependent upon a specific nucleus but only upon a specific molecular configuration within the molecule. From all the foregoing considerations it becomes clear that the growth process is due to a specific chemical reaction in which the auxins partake, so that the mental picture which we can make of the growth process is not that of a vague stimulation, but rather one in which growth has been brought back into the reality of a chemical reaction.

It is interesting to work out the simile of lock and key with these facts at hand. It will be clear that the activity of indole acetic acid is not due to some backdoor reaction, but it, like all active substances, is a key opening the same lock of growth. Thus the idea of the passkey is rehabilitated. The wards of the key which are essential for opening the mechanism of the lock are enumerated above, being double bond, carboxyl group, etc. But not every key having the proper wards will open the lock. Only if the proper grooves are present can the key be inserted into the lock. This comparison shows that both grooves (properties not essential for the reaction proper but determining the entrance of the substance) and wards (properties required for the specific reaction

under consideration) have to be considered in the analysis of physiological activity of these key substances for growth.

Some general conclusions can now be drawn which hold for specificity and activity of physiological substances both in plants and animals. When our experiments show deviations from a stoichiometric relationship between substances and a certain physiological process, this does not necessarily have theoretical significance. For it has been shown that in the case of auxin these deviations are due to a number of factors affecting the availability of auxin at its place of action, or which change the reactivity of the cell. Thus in comparing activities of physiologically active substances, we first must ascertain whether or not these substances actually reach the point where they must react, since innumerable secondary properties of the substances may affect their ability to penetrate the cell, move through the tissues, etc. If it is reasonably sure that this is the case, we must determine whether or not the applied substances affect the responsiveness of the cells to different extents. If these and other precautions have been taken, comparison of molar activities becomes significant.

In summarizing all the preceding facts, it can be said that apparently growth occurs as the result of certain chemical reactions, and we already know some of the substances involved. The final product, growth, cannot be defined chemically as yet; but in the near future it may be possible to build up a picture of growth and differentiation along purely physical and chemical lines, similar to the picture of respiration which has developed with amazing clearness in the last few years. At least one thing seems certain: the period when it was necessary to describe growth of plants as a stimulation lies behind us, and a much more concrete chemical concept is developing.

Thus far the auxins and their reactions have been investigated for their own sakes. But when enough is known about them they can be used as tools in a further investigation of processes connected with growth. One example has already been given in the case of the tropisms. Two more examples will follow showing

how general biological problems can be approached using the auxins not as problems but as tools.

In those cases where differences in growth and development are due to known hereditary factors, or genes, it is possible to analyze the action of such genes in terms of growth substances. In this way Van Overbeek studied a number of dwarf races of corn. [22] The difference between normal and dwarf varieties consists mainly in the fact that the stem does not elongate in the dwarf varieties so that all the leaves are placed close together in a sort of rosette. Van Overbeek found that, in general, auxin production, auxin transport, and sensitivity of the tissues for auxin were approximately the same in normal and dwarf corn plants. However, one outstanding difference between normals and dwarfs was found. In the dwarf plants the auxin produced in the stem tip was rapidly destroyed so that it did not reach the growing zones of the stems, whereas in the normal plants most of the auxin formed in the stem tip reached the growing regions. Thus the stems in the dwarfs could not elongate, and so the dwarfs were caused by excessive destruction of auxin. It was even possible to attribute the excessive auxin destruction to the high oxidative level of the tissues of the dwarfs, for catalase, oxidase, and peroxidase activity were much higher than in the normal plants. We can write the effect of the dwarf gene as follows:

Dwarf gene→	·→Increased oxidative level of cell	→excessive auxin destruction	→lack of stem elongation	→dwarf
geneticist→			←plant physiologist	

This shows how the plant physiologist approaches the action of the gene from one end, while the geneticist works from the other end. Some day we hope the scientist will be able to write the missing links in the reaction chain between the gene and its ultimate effect just as today the reaction chain for respiration between sugar and carbon dioxide can be written.

Another physiological problem of great interest is that of polarity. Both in plants and in animals it is evidenced by the formation of different organs at proximal and distal ends. Polarity in root formation, for instance, is very pronounced, and since we have just enumerated a number of factors involved in root formation it is worth while to ask where and how polarity comes into the picture. When the auxin transport inside living tissues was studied it was found that auxin moved only from the apex to the base, never in the opposite direction. [10] This means that the correlation phenomena affected by auxin can be transmitted only in one direction. In the case of root formation the auxin formed in the apical parts will ultimately collect at the base and thus cause basal root formation. Therefore the polarity of root formation can be attributed to the polar transport of a chemically known substance and thus the problem of morphological polarity is reduced to a physiological one. The next step obviously is to find out what physical forces determine this polar transport. So far the experiments have not given any answer to this question, but the problem is fundamental and warrants further thorough investigation.

In the preceding discussion two types of growth hormones have been discussed, the auxins and the root growth hormones, among which vitamin B_1 is the most important. There are many more hormones which unify the different parts and organs of the plant into an organism. It will not be possible to consider these other hormones in detail, partly because of lack of time and partly because they are insufficiently known. But a few should be mentioned so that it will be possible to get a composite picture of the plant as a whole, as something more than a simple colony of cells.

Young leaves develop at the expense of materials coming from the fully developed leaves. That these materials are more than simple carbohydrates has been known for some time. For instance a detached young leaf will show only slight growth in a solution containing sugars and the known growth substances. Re-

cently the chemical nature of some of the leaf growth factors which have to be added to the culture medium to make young leaves grow *in vitro* has been established by David Bonner and Haagen-Smit. [23] One of these is adenine, or a related purine derivative. This brings out once more the essential unity of plants and animals in that they require the same substances for vital functions, and it also shows how the utilization of sugar by different organs is limited by different chemicals acting as hormones. It is clear that in this way a differential development of the different organs is very easily accomplished by the simplest possible means.

Recently much experimental work has been carried out in connection with flower formation. It has been shown that the development of vegetative growing points into flower primordia is under the control of factors produced by leaves. [24] These factors can rightly be called flower-forming hormones and more evidence concerning their existence is continually accumulating. Determination of the chemical nature of these flowering hormones seems to lie well within practical possibility.

Finally, we may consider briefly another correlation carrier of which the nature is still completely unknown, but which connects the growing stem with the root system. It has been found that the function of the root system in promoting the growth of the whole plant consists of more than water and salt uptake. It is possible to separate the different root activities and to show that well-aërated roots contribute a factor *X* required for active stem elongation. This stem growth depends not only on the auxin coming from the stem tip, but also on the factor *X* coming from the root system.

If all these data are combined into a general picture, Figure 17 results. From this figure it is clear how all parts of the plant are interrelated by the formation of certain chemicals, hormones, in one part, which regulate the growth of other parts. Since all these relations are quantitative, an amazing unity within the plant is realized, so that one part cannot develop too much at the expense

of another. Only in this way is it possible for taxonomists to describe plants in terms of relative dimensions. If the leaves are small they will form little adenine and vitamin B_1, and therefore the new growth of young leaves and of the root system is limited. A small root system on the other hand will produce little of the factor *X* so that the elongation of stems is not only controlled by the root system but ultimately by the size of any existing leaves. Since the small leaves also produce little auxin their leaf stalks remain shorter than those of large leaves which produce more auxin. This picture fully brings out the original contention that a plant is a true democracy, "one for all and all for one." It also shows that the relations between cells and organs can be brought about by the simplest means. Certain cells lose the ability to synthesize one or another simple basic nucleus necessary for their build-up, and thus they become dependent upon other cells which still possess that ability.

So long as the intricate mechanism of the "plant democracy" was unknown, man had only one means to increase the yield of a plant product, such as fruit, wood, root, or flower. This was

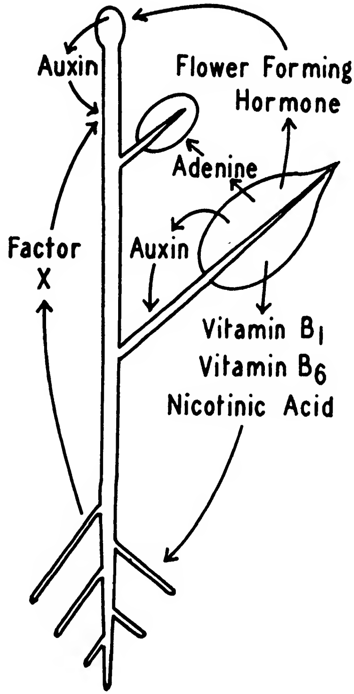


FIG. 17. Illustrating the interrelationships between the organs of a plant. Arrows indicate in which organ a specific substance is produced and the region to which it is translocated to perform its function. The sugars coming from the mature leaf are not shown. In downward sequence are indicated: apical bud, growing stem, growing leaf, mature stem, mature leaf, and root system.

by improving the growing conditions of the whole plant. But now new possibilities present themselves. If one knows exactly the factors which are required for growth of the specific organs in which we are interested, it may be possible to supply these artificially. The strength of the plant need not be equally distributed over all its organs; it may be possible to direct the development of the plant according to our special needs. The same effect is produced by plant breeding and selection: certain plants which are genetically unbalanced will have an excessive leaf production (for example, cabbage) or flower production (cauliflower), etc. By using plant hormones normal plants can be induced to produce the same unbalanced growth. And since the plant hormones are effective in minute amounts their practical use is possible also from an economic viewpoint.

The simplest example of unbalanced growth due to artificial application of growth factors is the root growth in vitro. Here the presence of the rest of the plant is superfluous. Since it is necessary to supply sugars in addition to vitamin B_1 and other root-growth hormones, off hand it does not seem practical to produce carrots or beets by cultivating root tips in vitro. But in other cases unbalanced growth can be used commercially.

The first example of this is the rooting of cuttings. Normally a leafy shoot of a plant produces root-forming factors, which ultimately will reach the root system and help build it up. If the shoot is cut off and made into a cutting, this production of hormones continues and they accumulate near the basal cut surface, which often results in differentiation of roots, and rooting of the cutting. But in a number of instances the hormone production is inadequate to initiate roots at the cut surface, and the cutting will die without rooting. In a remarkable number of such cases the application of auxin is effective and will result in a successful rooting of the cuttings. This method is now being used commercially by many large nurseries for their hard-to-root cuttings. In some cases the root production on cuttings treated with auxins seems completely disproportionate to their size. This is not a

disadvantage, for usually such cuttings will develop into normal plants, often at a faster rate than cuttings with only a few roots.

In other cases it has been reported that the set of fruit can be considerably improved by spraying trees previous to a period of excessive fruit drop soon after flowering. Here we stand only at the beginning of interesting developments.

Improved germination, better seedling growth, more abundant flowering are all effects claimed to result from treatment with small amounts of auxins.

A recent practical application of plant hormones which has been widely publicized is that of vitamin B_1. It is known that most plants form this substance in their leaves in the light, and that the root system needs it for continued growth. Some plants have an insufficient production of vitamin B_1 to support themselves, and therefore are dependent upon some outside source. In general these plants (such as camellias, azaleas, and other humus plants) receive their B_1 from the soil, which means that they have to be grown in a medium containing or producing vitamin B_1. A good garden soil, manure, peat, all contain some B_1, and therefore humus plants can be grown in such soils. But many soils are very low in organic matter, and therefore they can support only plants with a sufficient vitamin B_1 production of their own (such as most crop plants). Experiments have shown that root development is sometimes subnormal. In those cases it is possible to increase the root system significantly by watering with very dilute solutions of vitamin B_1 (e. g., once per week with a 1:100,000,000 solution). This in turn will result in better growth of the whole plant.

Work under way with leaf-growth hormones opens new possibilities for plant engineering, and other opportunities will appear when flower-forming substances can be synthesized and introduced in plants.

All these present and future practical applications of plant hormones are based exclusively on systematic research which was undertaken to find out more about the principles of plant growth.

But since so much of our horticulture and agriculture depends on growth, any increase in our theoretical knowledge must lead to new or improved methods of growing plants. Such new discoveries are in no way mysterious. If we happened to stumble upon them, or if by accident a substance was found which increased plant growth, then one might speak of mysterious substances. But there is nothing of this nature about auxins or vitamin B_1, and there was little left to accident in their discovery as plant hormones. If this situation is compared with the significant advances in chemical and electrical engineering, then it is seen that those advances too were a result of better understanding of the basic chemical and physical forces and laws. This is quite logical. Most of the simpler practices, which can be discovered by accident, have already been found in the bygone centuries. It becomes more and more improbable that someone will stumble by accident upon something completely new. But on the other hand the ever-increasing body of scientific knowledge forms a more and more certain basis for new developments. Whereas this has been recognized for a long time in the case of chemistry and physics, the importance of basic knowledge for improvement in agriculture and horticulture is only slowly becoming realized outside academic circles. But with the accumulation of more basic data concerning the growth and development of plants a new science, plant engineering, is slowly emerging. This can contribute materially to a better and more logical utilization of the possibilities which the life activities of plants offer for our human well-being.

III

EXPERIMENTAL STUDIES ON THE FUNCTIONS OF THE FRONTAL LOBES IN MONKEYS, CHIMPANZEES, AND MAN

By J. F. Fulton

School of Medicine, Yale University

It became obvious, at the end of the last century, through a series of preliminary studies,[1] that much could be learned concerning the activities of the human nervous system by a comparative approach, i. e., by the study of the brains of man's nearest of kin in the animal scale, the primates.[2] Their brains could be exposed and stimulated, homologous areas could be removed, and if such procedures were carried out in anthropoid apes, information could be obtained that was highly relevant to man. Owing, however, to difficulty in obtaining experimental subjects, the comparative approach was not actively followed, save for the stimulation experiments begun in 1901 by Grünbaum and Sherrington, which were continued for a period of several years and ultimately published by Leyton and Sherrington in 1917 under the title "Observations on the Excitable Cortex of Chimpanzee, Orang and Gorilla." In 1930 opportunity presented itself to make a systematic comparison of the responses to stimulation and ablation in a large series of primate forms, and the summary which will follow is based largely upon the continuation of these studies, first summarized with Keller in 1932.

1. Carried out by David Ferrier (1876), and by Beevor and Horsley (1890), Schäfer (1888, 1898), and finally by Sherrington (1892). See page 288 ff.

2. See Appendix, page 73.

ORGANIZATION OF THE NERVOUS SYSTEM

The central nervous system is a vast assemblage of cellular units, the neurons, which serve to adjust the individual organism to changes which occur in its immediate environment. The elementary unit of reaction is the reflex arc. It consists of a *receptor* cell (sensory neuron) and an *effector* cell (motor neuron). Lying between the two in the central nervous system there are gen-

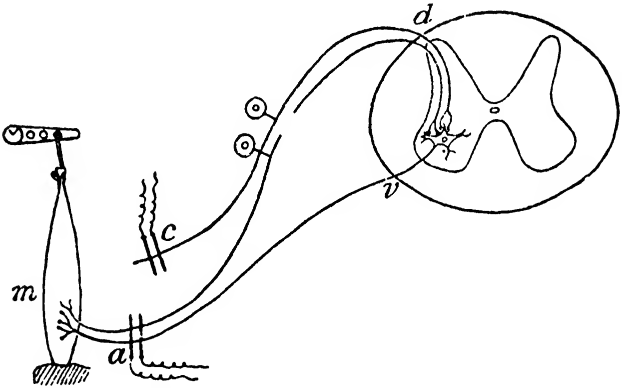


FIG. 18. Diagram of muscle-nerve preparation to show paths of nerve impulses in reflex arc. Right: cross section of spinal cord with spinal nerve extending to muscle tissue (*m*). *a*, *c*, electrodes for stimulating nerve fibers; *d*, dorsal root of spinal nerve which carries sensory impulses to spinal cord; *v*, ventral root of spinal nerve which carries motor impulses to muscle (*m*).

erally one or more *adjustor* neurons, sometimes called interneurons or internuncials. In the reflex arcs, time elapses between the arrival of the impulse from the receptor and the initiation of the executant impulse in the effector. This time-lag is referred to as "central reflex time," and its magnitude varies with the number of interneurons encountered in the reflex circuit (Fig. 18).

After this elementary statement of the reflex concept, it is possible to take up a few details concerning the organization of

the nervous system itself. We must first consider the concept of levels of function.

Levels of Function

In man and vertebrate animals the spinal cord constitutes the basic level of functional activity, and its neurons are organized to make possible certain patterns of reflex response. Perhaps the best known is the flexion reflex: if the toe of a spinal cat is pinched, the limb withdraws. One can read into this reaction an obvious element of purpose, i. e., it is designed to move the limb out of harm's way. The reaction, however, involves not a single muscle but a group of muscles—some twenty or thirty in fact—and it is therefore a complex affair.[3] This leads to the first broad generalization about the nervous system, namely, that *it is organized functionally to produce movement—it "thinks," so to speak, in terms of movement, rather than in terms of muscles.*[4]

There are many other patterns of response observable in spinal animals such as reflex emptying of the bladder, reflex sweating, patterns of sexual behavior, etc.; but viewed broadly, these reactions are all stereotyped, and they all have a similar latency of response. Relatively they are actually far more simple than the reflex adjustments possible in intact animals.

Another important level of function lies just anterior to the spinal cord in the *medulla oblongata* (Fig. 19). Here, more complex movement patterns are organized, such as those coming from the labyrinth, the postural reflexes arising in the neck muscles, swallowing, and regulation of the level of blood pressure. The next higher level of importance is the *hypothalamus* which plays an important role in regulation of visceral functions.[5] Higher still is the *striatum* concerned again in postural adjustments, and finally at the highest level of all are the cerebral *hemispheres* which are responsible, among other things, for regu-

3. Sherrington, 1910.
4. Sherrington, 1931, p. 21.
5. Fulton, et al., 1940.

lating those intellectual functions which distinguish man from beast.

Dominance of the Anterior Segments. The primitive nervous system, e. g., that of the earthworm, is organized in terms of

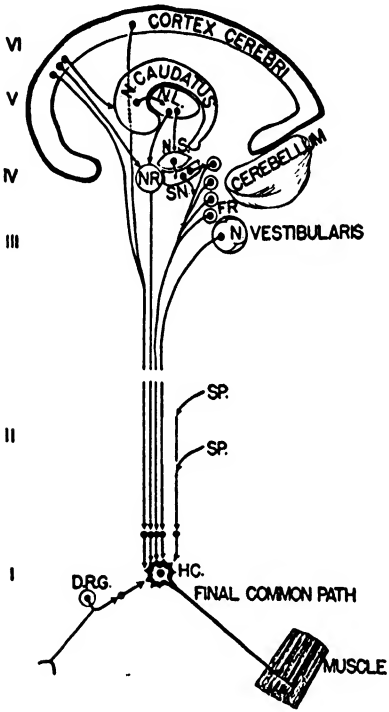


FIG. 19. Diagram showing paths of nerve impulses in central nervous system. I, neuro-muscular; II, spinal; III, hind-brain; IV, mid-brain; V, striatum; VI, cerebral cortex, neurons in areas 4 and 6 only are shown. N.L. = nucleus lenticularis. N.S. = nucleus subthalamicus. D.R.G. = dorsal root ganglion. H.C. = ventral horn cell. (From Cobb, S., *A preface to nervous disease*, 1936.)

levels, and it is obvious even in the lowest forms that the more headward levels of function have, in the course of evolutionary development, become dominant over the more caudal segments.

The extent of this dominance, however, varies widely in different species, and it is for this reason that much importance is attached to *comparative* studies of central nervous function.

An illustration of differing degrees of cortical dominance is found by comparison of cat and monkey. The cat (or dog) after removal of its cerebral hemispheres exhibits, after a time, essentially normal locomotor movements: it is able to walk and to maintain itself upright against the force of gravity. A monkey, however, when it loses its cerebral hemispheres is never able to walk again because in these primate forms the control of locomotor movement has, in the course of evolution, become almost completely taken over by the cerebral cortex.[6] Locomotor function has become, as we say "encephalized."

Such comparative studies have progressed in two important directions: (*a*) analyses of the gradual *encephalization* of function; and (*b*) investigation in a given species of functional localization in various parts of the hemisphere. The title of this chapter should really have been "encephalization and functional localization in the primate frontal lobes." However, I feared that this might frighten the physical scientists, for although they invent words far more terrifying, they always seem a little intimidated by the simple words which we devise in the biological sciences!

ENCEPHALIZATION OF FUNCTION IN THE PRIMATE FRONTAL LOBE

There are certain relatively constant features of organization of the brain in all primate forms. The first is the existence of a fissure, usually referred to as the "central sulcus," which separates the rostral from the caudal half of the hemisphere. Even in the lowly tarsier a small depression may be found in the otherwise smooth surface of the cerebral mantle which represents the homologue of a central sulcus. The posterior half of the cerebral hemisphere, i.e., the part lying caudal to the central sulcus, is

6. Bieber and Fulton, 1938.

primarily sensory in function and receives the great sensory projections from the thalamus and from lower spinal levels. The frontal lobes constitute the anterior half of the hemispheres and include all tissue lying rostral to the central sulcus (Fig. 20). This half of the hemisphere is primarily motor in function, but it also has a vast assemblage of interneurons which increase enormously in extent as one ascends the primate scale. The front part of the frontal lobe is made up entirely of those interneurons; the region itself is referred to as the "frontal association areas,"

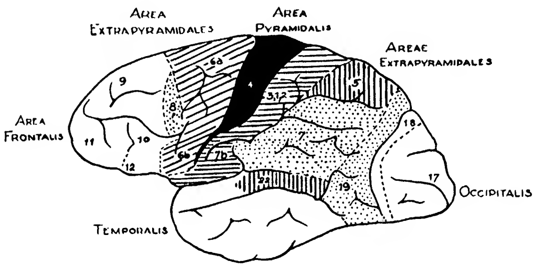


FIG. 20. Diagram of the left cerebral hemisphere of the Chimpanzee, a Primate.

and neurologists and our colleagues in physiological psychology would have us believe that these association areas are primarily concerned with higher intellectual functions, such as memory and the capacity to plan for future action.[7]

How do we know that the brains of chimpanzee and man are more encephalized than those of pithecoid monkeys, lemuroids, and tarsiers? A suggestion comes from the study of stimulation of the frontal lobes. I have never stimulated the tarsier—the animal is too rare and too precious—but I have stimulated pottos, galagos, and marmosets which belong to the primitive group (Fig. 21). The Germans call them appropriately *Halb-affen*—half monkeys. It is strikingly true that only the most stereotyped movement patterns can be evoked from the cerebral cortex of

7. See Hilgard and Marquis, 1940.

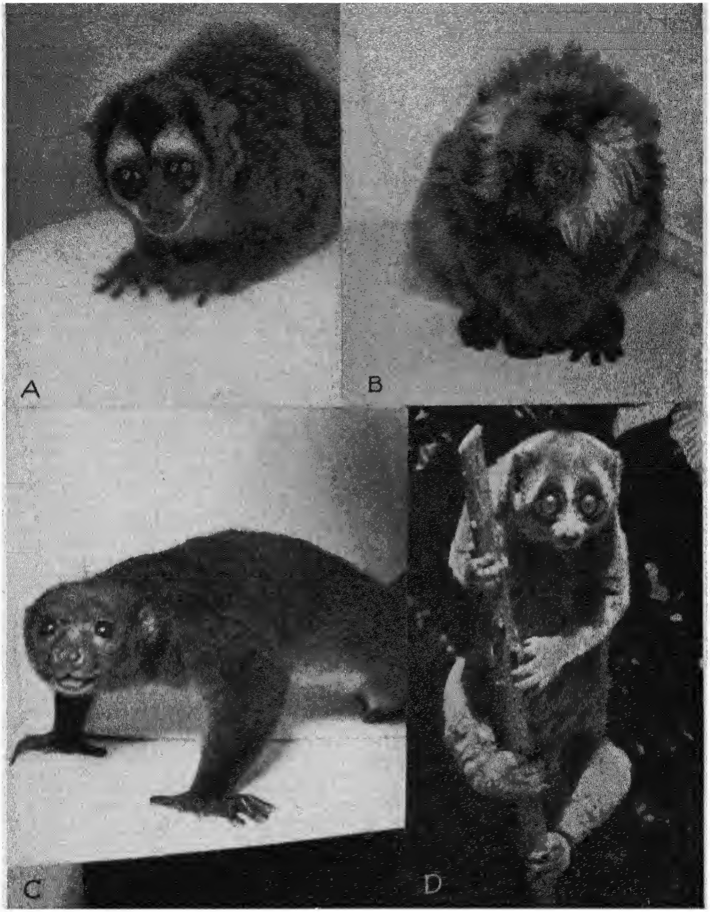


FIG. 21. Four primitive monkeys: A, Marmoset; B, Madagascar Lemur; C, Potto; D, *Nictocebus loris.*

these primitive primates, whereas in the chimpanzee and man highly discrete movements of individual fingers, corner of the mouth, the eyelids, the vocal cords, and other highly selective

movement patterns can be evoked by cortical stimulation. Moreover in man and chimpanzee there is a mosaic of sharply circumscribed foci controlling these specialized movements (Fig. 22); in the lower primate forms the distribution of excitable points is homologous, but excitable foci are less individualized.

More convincing evidence of encephalization comes from study of the effects of isolated removal of specific movement areas in the frontal lobe. If the leg area is removed, the leg itself becomes

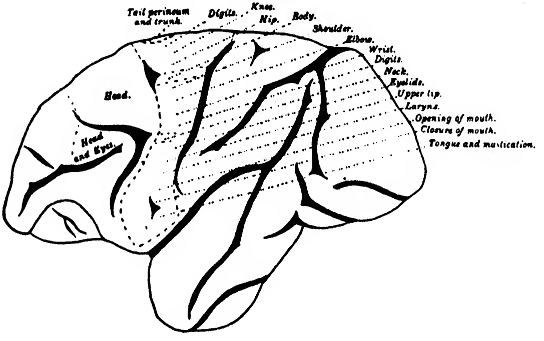


FIG. 22. An early diagram of the motor areas of cerebral cortex. (After Jolly and Southerland.)

paralyzed. *The extent and duration of that paralysis, however, varies widely in different forms.* It was noted above that a cat, after complete removal of the hemispheres, is able, after a brief period of weakness, to walk away as if nothing much had happened. The marmoset, galago, and potto behave much as would a cat, although they exhibit a somewhat more enduring paresis. Thus on removal of the leg area of a galago or potto, the leg shows obvious weakness for a few days, but power is later regained, and the animal ultimately moves about as if nothing had happened. Pithecoid monkeys, such as the macaque, exhibit a profound paralysis when the foot area is removed, but recovery gradually occurs and after a few months the animal may move its

foot almost normally in ordinary locomotor movements; signs of weakness, however, persist indefinitely following fatigue. Even six months or a year after destruction of the leg area the animal will begin to drag its affected foot following a fatiguing chase. Hence integrative foot movements are sufficiently encephalized in the macaque monkey to give some degree of permanent deficit when the foot area is removed. Turning to chimpanzee and man, a far more profound paralysis results when the foot area is ablated; both exhibit a permanent and readily recognized motor deficit when the foot area of the frontal lobe is destroyed.

If one compares the *hand* instead of the foot the difference is even more striking, for a human being whose arm area has been destroyed can never again use his fingers for the fine movements essential to play the piano, or even to button a shirt. Monkeys, however, although they do not ordinarily play the piano, exhibit other exquisitely delicate movements of the fingers, and these show far less ultimate impairment following a lesion of the arm area than those of chimpanzee and man. Hence we are led to conclude that the patterns of skilled movements are more highly encephalized in man and ape than in monkey.

Many parallels can be given in the sensory sphere of corresponding encephalization of function—to take one example, *vision*. In fish the optic system has no significant connection with the forebrain, and removal of the cerebral hemisphere causes no detectable impairment of vision. According to Schrader (1889), the same is true of reptiles and birds. However in rodents, particularly rabbit and rat, the visual cortex becomes encephalized; if the occipital lobes are removed object vision is impaired but the capacity to discriminate between differing intensities of light still remains unaffected.[8] Marquis (1932, 1938) has shown that dogs appear for a time completely blind after removal of the occipital lobes, but that brightness discrimination eventually returns with slight impairment and object vision is entirely de-

8. Lashley, 1932.

stroyed. In monkeys object vision and brightness discrimination disappear entirely and only a modicum of brightness ever returns.[9] Man is rendered completely and permanently blind by a bullet through the occipital lobes, and he fails to regain any consciousness whatsoever of light when this part of his cerebral hemisphere has been destroyed.

Therefore when one approaches a given function of the human brain from the comparative standpoint by studying a particular function in a series of animals extending from tarsier and the lemurs up through the monkeys and the great apes, one can "extrapolate" quite accurately and predict how man will differ in regard to this particular function.

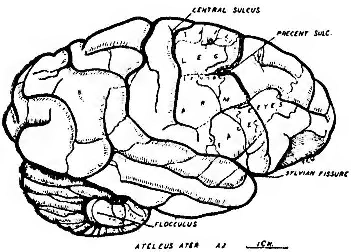


FIG. 23. Diagram of right cerebral hemisphere of the tailed monkey (*Ateleus ater*) showing the localization of the tail, leg, arm, and various other areas. (After Fulton and Dusser de Barenne.)

In passing it is perhaps interesting to know that control of movements of the tail in the prehensile-tailed South American monkeys is so completely encephalized that when the tail areas are removed bilaterally, the tail itself becomes permanently and completely paralyzed, exhibiting only a spastic prehensility similar to the condition encountered among primates after decerebration (Fig. 23).[10]

So much then for the principle of encephalization. Clinical neurologists have a tiresome way of telling physiologists, "Well, you are working on cats and monkeys and your results are, of course, not applicable to a human being." My reply to this is that if one approaches human function from a comparative standpoint, far more basic light can be thrown upon it than by all the studies in Christendom carried out on man himself without reference to his forebears in the evolutionary scale. A study of the

9. Marquis and Hilgard, 1937. 10. Fulton and Dusser de Barenne, 1933.

monkey is quite meaningless so far as the rest of the animal kingdom is concerned, unless study of a given function is compared with the corresponding studies on forms occupying differing positions in the animal scale. The comparative approach to the human brain tells us why it works as it does.

FUNCTIONAL LOCALIZATION IN THE FRONTAL LOBES

The greater part of the cerebral cortex is made up of six primary layers of cells, but each layer has variable characteristics and many analyses have been made of the *cellular architecture* of the brain. On the basis of such studies, the frontal lobes have come to be divided into a number of structurally discrete fields to which (for reasons best known to himself) Korbinian Brodmann (1909), the Armenian histologist, gave an erratic series of numerical designations. Thus, the motor area—which on stimulation causes movement of skeletal muscles—is known as area 4; the adjacent premotor region as area 6; another area which on stimulation moves the eyes is numbered 8, etc. The map in Figure 24 gives these numerical designations as modified by Dusser de Barenne and McCulloch, in 1938; the capital letters L, F, and A denote respectively leg, face, and arm, so that area L.4 is the leg area and A.6 the part of the premotor area concerned with the arm, etc.

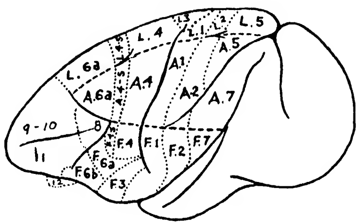


FIG. 24. Map of the frontal lobes of cerebral cortex of a pithecoid monkey. (After Dusser de Barenne.)

Motor Fields (Areas 4 and 6)

The excitable characteristics of these areas have already been described, *i. e.*, a mosaic of discrete foci can readily be demonstrated. When they are removed from monkeys or chimpanzees

an enduring paralysis occurs; indeed, if areas 4 and 6 (including leg, arm, and face) are destroyed bilaterally in an adult monkey, the animal is as gravely paralyzed as when both cerebral hemispheres are removed. This not only means that the motor functions are encephalized to this surprising extent in higher primate forms, but it also indicates that they are *localized* in a definite, rather small area within the cerebral hemisphere.

This general statement must at once be qualified, and the qualification is as significant scientifically as the generalization itself. The reservation has to do with young animals. You are all familiar with the biological concept that ontogeny tends to repeat phylogeny, that is, embryological development is a recapitulation of evolutionary history. It is undoubtedly a fact of prime importance that when the motor and premotor areas are removed from monkeys shortly after birth, as has been done in the recent studies of Dr. Margaret Kennard (1938), the baby monkeys behave pretty much as do adult galagos and pottos after a similar lesion: they have only slight motor paralysis. However as these baby macaques develop, signs of deficit begin to appear, but the defects never become as grave as the paralyses which follow a corresponding ablation in an adult animal. *There exists in the infant nervous system of a pithecoid monkey greater capacity to reorganize its functional localization than in the adult,* and the pithecoid baby behaves as if it were an adult lemuroid.

The Frontal Eye Fields (Area 8)

The frontal eye fields have engaged the attention of experimental neurologists since their existence was first clearly demonstrated at the end of the last century,[11] but their functions are seldom discussed in considering the frontal lobes, and no one had attempted to study the effects of isolated ablation prior to the work of Kennard and Ectors.[12] They also studied the excitability of the cortical eye fields, and confirmed earlier work suggesting

11. Beevor and Horsley, 1890.
12. Kennard and Ectors, 1938; Kennard, 1939.

that stimulation causes: (*a*) conjugate deviation of the eyes toward the side opposite with a tendency toward turning the head in the same direction when the stimulus is strong; (*b*) pupillary dilatation; (*c*) opening or closing of the lids (different foci), and (*d*) lachrymation of the eye on the opposite side. There is thus evidence of focal representation in area 8, but the foci are less discrete than in area 4.

Isolated removal of area 8 causes a conspicuous group of symptoms, or syndrome, which is well marked and easily recognizable; the operation has important localizing value in clinical neurology. Unilateral ablation of area 8 causes conjugate deviation of the eyes toward the side of the lesion, forced rotation of the body toward that side during locomotor movements, failure to heed objects on the opposite homonymous side—"pseudo-hemianopia"; all these effects are transient, disappearing entirely within two to six weeks.

Removal of no other part of the cerebral hemispheres causes a similar syndrome. Thus ablation of the occipital lobe gives an enduring hemianopia but causes no forced rotation. Removal of areas 4, 6, 3-1-2, 5, or 9-10-11-12 produces neither rotation nor visual disturbance. Ablation of the angular gyrus region, however, gives a curious turning movement. The animal sits and rotates as though impaled upon a corn stalk. The disturbance is probably vestibular.

When area 8 is removed from both sides, the animal appears dazed, fails to respond normally to objects brought into its field of vision, and its facies become immobile as in Parkinsonism. The gaze is "fixed" for the first week or ten days, but thereafter the animal becomes restless, distractable, and shows permanent motor hyperactivity so long as it is in a lighted room. When the room is darkened the motor activity ceases.[13]

The Face Areas

The face areas have been little studied since the early work of the Vogts. Dusser de Barenne, McCulloch, and Ogawa have investigated the effect of local strychninization of the subdivi-

13. Kennard and Malmo, unpublished.

sions of the face area indicated in Figure 24. Such activation of any part of the face areas failed to affect the arm or leg areas, save for F.4-*s* which caused slight suppression of electrical activity in A.4 and L.4. Strychnine applied to F.4-*s* and to F.1 (postcentral) suppressed electrical activity of F.4, indicating that F.4 is normally regulated by adjacent regions. Walker and Green (1938), investigating the effects of stimulation of the face areas, have largely confirmed the work of the Vogts in finding that area F.4 caused brief contractions of individual muscle groups, F.6*a* yielding more complex movements of the premotor type, F.6*b* masticatory and deglutatory movements and also reactions of the vocal cords, salivation, etc. They also noted, as had the Vogts, that stimulation of the lower part of F.6*b* caused inhibition of respiration; but this finding is probably to be reinterpreted in the light of Bailey and Sweet's recent (1940) disclosures concerning respiratory representation on the orbital surface of the frontal lobe.

Analysis of the effects of ablating the face areas led Green and Walker to conclude that area 4 sends direct cortical projections to the final motor neurons of the labial muscles, the muscles of the food pouches, and muscles of the tongue. The premotor area (6*a*) probably has few direct motor projections, while F.6*b* possesses independent projections which are mainly concerned with gross movements of the tongue, fauces, and vocal cords. The lower facial musculature is represented almost solely on the contralateral cortex, while the upper facial and pharyngeal and laryngeal have a larger bilateral representation.

Boernstein (1940) has recently proved that taste perception is localized in areas F.3, 4, and 6, since after bilateral destruction of these areas a monkey will drink bitter water as rapidly as sweetened water.

The Respiratory Area

The striking results of Pitts and his collaborators (1939) on the localization of the respiratory mechanism of the medulla have

aroused fresh interest in the regulation of respiratory movements. They have been able with the Horsley-Clarke technique to isolate the ventral reticular formation of the medulla (near the inferior olive), an *inspiratory area.* When this region is stimulated electrically, the chest and diaphragm remain in fixed maximal inspiration, rhythmic breathing is abolished, and death may occur if the stimulation is continued.

An *expiratory area* was similarly delimited and was found to lie "dorsel to, slightly cephalic to, and cupped over the cephalic end of the inspiratory reticular formation." The expiratory movement evoked by stimulating this area is somewhat less intense than the inspiration evoked from the inspiratory area, but it involves both thorax and diaphragm and in some instances is maximal. If stimulation persists, a maximal expiratory posture may be maintained for two to three minutes, but death does not supervene, since inspiration gradually is resumed. The respiratory movements observed in these studies were well-integrated respiratory acts involving simultaneously thoracic and diaphragmatic musculature and hence cannot be due to stimulation of efferent or afferent fiber tracts, but rather to a focus of neuronal integration.

Respiration is also influenced from supramedullary levels. Thus, Magoun (1938) finds that the well-known inhibitory effect on respiratory movements resulting from hypothalamic stimulation is considerably depressed when the hypothalamus is stimulated some weeks after removal of the cerebral cortex. This suggests that the normal response is due to activation of corticofugal fibers which pass from the cortex through the hypothalamic area.

New light upon the role the cerebral cortex plays in regulation of respiratory movements has recently come from the experiments of Bailey and Bremer, who found that, when the central end of the vagus nerve was stimulated, conspicuous changes occurred in the electrocorticogram of the orbital surface of the frontal lobe in cats and in no other region. More recently, Bailey

and Sweet (1940) find, too, that electrical stimulation of the orbital part of the frontal lobe causes inhibition of respiratory movements. It had formerly been thought[14] that area F.6*b* (Fig. 24) on the lateral surface of the frontal lobe also inhibited respiration. Bailey believes that effects when obtained from this area are probably due to spread of current along the moist orbital bone to the orbital surface of the frontal lobes. Acceleratory effects on respiration may be obtained from points on the lateral surface of the frontal lobe—in the monkey from the anterior part of area L.6, near the midline (Smith).

Frontal Association Areas

The influence of specific ablations of the frontal lobes upon intellectual processes has been studied at length and recently reviewed by Jacobsen (1939).[15] The character of the intellectual deficit which follows bilateral ablation of the areas 9–10–11–12 of Brodmann has been described in objective terms by a number of simple procedures, such as the delayed reaction test and the stick-and-platform problem, both of which involve temporal behavior, and both of which are disturbed following bilateral ablation of the frontal areas in monkeys, chimpanzees, and man. After such lesions, the experimental subject tends to become restless and distractable; if area 8 is included the restless hypermotility becomes even more pronounced. The animals continue to exhibit a lively interest in their immediate surroundings. Jacobsen also noted that chimpanzees after such ablations were seemingly immune to experimental neuroses, being quite undisturbed by failures in discrimination tests.

On the basis of Jacobsen's experimental findings, Moniz introduced the procedure which has been much discussed as a form of treatment for certain psychiatric derangements in man. Instead of doing a "Dandy"—as one of my neurosurgical friends terms bilateral frontal area removal—Moniz merely interrupts the

14. See Bucy and Case, 1936; Walker and Green, 1938; Smith, 1938.
15. See also Finan, 1939.

projection fibers from both frontal association areas with a leucotome through a pair of burr holes. Dr. Lyerly of Jacksonville, Florida, accomplishes the same thing by an obstetrical approach —or perhaps I should say "technique"—for in place of a leucotome, he inserts a virgin speculum through the burr holes into the frontal lobes and severs the frontal projections under direct observation. But whether one approaches the anatomical problem obstetrically, neurosurgically, or with a tank gun,[16] many institutional cases regarded as hopeless have as a result of interruption of their frontal projection become euphoric, cheerful, and able to cope with the exigencies of existence. The effects of bilateral frontal lobotomy are now under close scrutiny and are being compared with the results of other procedures such as insulin and metrazol convulsions. It is too soon to pass judgment upon the relative merits or justifiability of any of the three forms of therapeutic treatment, but the lively interest which these new procedures have aroused indicates the close correlation which is now coming about between experimental physiology of the nervous system and clinical neurology.

DISCUSSION

Finally, I should like to say a few words about the general functions of the frontal lobes. The first point emerging from recent studies is a purely practical one and has to do with the localizing value of specific symptom complexes to the clinical neurologist. Head turning toward the side of the lesion in frontal lobe cases is a common phenomenon, but is seldom commented upon in neurological literature. Similarly, the conjugate deviation of the eyes to the side of the lesion, so often seen after craniotomies, is generally thought due to involvement of the occipital lobe. Actually, complete removal of the occipital lobe does not of itself induce conjugate deviation of the eyes or head turning or forced rotation. The symptom in our experience is referable

16. See Freeman and Watts, 1938; and, more recently, Rylander of Stockholm, 1939.

to some small area, namely, the eye fields, when it is present as an isolated symptom; or in association with other disturbances it points to involvement of area 8.

The second point is more general. I have just finished editing and marking for press thirty-four papers on the hypothalamus.[17] Those who heard or will read these papers may conclude that all visceral organs and processes have some degree of representation in the hypothalamus. In a measure this is true; but I never cease to marvel at the perspicacity of old Hughlings Jackson, who insisted that everything which is represented at one level in the nervous system is *re*-represented at higher levels, and this general concept accounts for much of the apparent multiplicity of function now assigned to the frontal lobes. I have not referred to the vasomotor, sudomotor, glandular, or gastro-intestinal disturbances which result from stimulation or ablation of the frontal lobes; but mention has been made of the fact that lachrymation may be induced from area 8, salivation from areas F.4 and F.6*a*, and pupillary effects were likewise mentioned. All these reactions can also readily be obtained from stimulation of specific areas in the gray matter of the hypothalamus, even after the frontal lobes have been removed, or after both cerebral hemispheres have been ablated and enough time has elapsed for the projection fibers to degenerate. But the adjustments which occur in the hypothalamic animal, notably those in the sphere of temperature regulation, in regulation of blood pressure, etc., are less adequate than when the higher integrative levels of the frontal lobes continue intact.

One must remember, furthermore, that disturbances of function may occur at any level and that they may involve nearly any function, often in an isolated manner. Hence it is a matter of some significance that immediately after section of the frontal projections in conscious human beings Watts and Freeman report that the blood pressure drops, the skin of the extremities and face becomes warm and dry; and Lyerly, confirming these

17. Fulton, Ranson, and Frantz, 1940.

observations, also points out that the pupils tend to become more constricted and a moderate degree of enophthalmos is likely to develop. These changes, moreover, are enduring. With them occur profound alterations in the psychomotor status of the individual; in some instances, there have been striking disappearances of phobias and anxieties previously present.

I have no comments to make, however, on the justifiability of the therapeutic procedure as a procedure. This is a matter for experienced clinical observers to decide. But from the physiological standpoint I find deep significance attaching to the concomitant alteration in the autonomic and psychomotor spheres which accompanies interruption of the frontal projection fibers. It lends further support to the belief, well substantiated on other grounds, that the frontal lobes, which form the principal motor division of the cerebral cortex, are in command of mechanisms affecting not only the somatic but all phases of autonomic regulation, and that the same functional organization which makes this possible also determines in large measure the behavioral and psychomotor status of the individual. Lesions posterior to the Rolandic fissure give specific sensory disturbances without conspicuous psychomotor disturbance, save possibly when the lesions involve the temporal lobes. Hence, in the comparative study of the frontal lobes and their association with lower functional levels, particularly the hypothalamus, one finds a challenging approach to many of the most fundamental problems of neurology and psychiatry.

APPENDIX

THE PRIMATE GROUP

In passing, it must be recalled that the *Primata*, the first order of mammals, was recognized by the great Swedish naturalist, Karl von Linné, in the earliest edition of his *Systema natura* (1735), but the name "primata" was not actually adopted until the celebrated tenth edition of his book had appeared in 1758. Placing great emphasis upon the

teeth as the basis for classification, Linné originally described a primate as "a quadraped with four parallel incisors, single canines, two pectoral mammae, hands (rather than paws), two complete clavicles and an arboreal habitat." In defining the primates today one adds to these characteristics: five fingers and five toes, usually with flat nails, the first and second digits being generally opposable, but the thumb and hallux may be atrophic. Usually a single offspring is produced at birth in a completely helpless condition; twinning occurs normally in a few of the lower primates and multiple births may occur rarely in all primate forms including man. The bat, originally included by Linné among the order of primates, is now assigned to a separate order.

There are three suborders of primates: *Tarsioidea*, *Lemuroidea*, and *Pithecoidea*.

Tarsioidea. The group of animals nearest the base of the primate tree is believed to be the tarsioids which are represented by only one existing genus, *Tarsius* Storr (Fig. 25). The recent studies of Matthew, Chester Stock,[18] and others indicate that tarsioids similar to the existing species were extant in the early Eocene, and Stock has found that their distribution in the Eocene extended to this hemisphere. At present they occur only in a limited part of Melanesia, namely, the southern Philippines, the Malay peninsula, Celebes, northern Borneo, and possibly in some intermediate islands. Few if any other mammals have survived from the early Eocene and hence the Tarsius must be regarded as among the oldest surviving mammals.

Interest in the comparative study of cerebral function prompted a trip to the southern Philippines two years ago to study the existing species with a view to transporting them to this country by air. The tarsiers had not hitherto been brought either to America or to Europe, although the attempt had been made on several previous occasions. Too little was known of their feeding habits, and of their uncertain temperament. While in the Philippines I made the acquaintance of a retired army captain, Mr. Norman Cook, who, while stationed in the Moro Gulf in southern Mindanao, had succeeded in keeping tarsiers as pets; and he passed on to me his notes concerning their feeding habits and other features of their natural history.[19] Profiting by this information, I succeeded in keeping a number of specimens alive and transported one by

18. See Matthew, 1928; Stock, 1933, 1934, 1938.
19. See Cook, 1939; also Le Gros Clark, 1924.

FIG. 25. A specimen of *Tarsius* from the Philippines.

clipper plane back to Honolulu; in the early clippers odds were against this, because livestock had to be carried in an unheated baggage compartment which at 12,000 feet frequently fell to a subfreezing temperature. I made arrangements, however, with an enthusiastic young naturalist, John Eckmann, to bring back some specimens by boat and these

arrived in November, 1938. A pair, male and female, have survived for eighteen months in my laboratory.[20] The male is just becoming sexually mature; the female, whose sex cycles have been closely studied, will, we hope, sometime reproduce.

The physical and structural characteristics of the tarsioids indicate that they stand in the region of transition between the insectivores and the primates. Hence, a study of their brains from the structural as well as from the functional standpoint is of the greatest possible interest.[21]

Lemuroidea. An offshoot of the primate stem is found in the lemurs, galagos, pottos, aye-ayes, and other little-known animals of Central Africa and Madagascar which answer to Linné's definition of a primate. They are less significant from the point of view of primate evolution than are the tarsioids, because they probably represent an offshoot considerably removed from the base of the primate tree. However, they are more available than are the tarsioids and exist in many species.

Pithecoidea. The greatest number of primates is grouped under the pithecoids, which include two large divisions of the Old World and the New World monkeys, all the apes, and man. Hence from the point of view of comparative study of the brain, the suborder *Pithecoidea* includes those forms to which attention is ordinarily directed. The general relationships of the primates are indicated in the following table.

TABLE I. PRIMATE CLASSIFICATION

Order I: Primates

SUBORDERS

I. TARSIOIDEA, a rare, insect-eating primate which has changed little since Eocene times and exists in only one Genus *Tarsius* Storr (several species or varieties). They are nearer the base of the primate stem than any other extant form.

II. LEMUROIDEA, a lowly offshoot of the primate stem, with tarsioid affinities, exist in four families:

1. *Lemuridae*	Lemur
2. *Lorisidae*	Loris, Potto
3. *Galagidae*	Galago ("Bush-baby")
4. *Daubentoniidae*	"Aye-aye"

20. Catchpole and Fulton, 1939.
21. See Woollard, 1935; also Woodward, et al., 1919, and Le Gros Clark, 1934.

III. PITHECOIDEA, an enormous group including Old World (*Catarrhini*) and New World (*Platyrrhini*) monkeys, the anthropoid apes, and man, divided in six families

Monkeys
(*Platyrrhini*)

1. *Hapalidae*	Marmoset
2. *Cebidae*	Ring-tail, Spider, Woolly, Howler

(*Catarrhini*)

3. *Cercopithecidae*	Langur, Four-fingered, Green, and Red Military Monkey, Rhesus Macaque, Japanese Ape, Baboon, Mandrill

Anthropoid Apes and Man
(*Catarrhini*)

4. *Hylobatidae*	Gibbon, Siamang
5. *Pongidae*	Orang-utan, Chimpanzee, Gorilla
6. *Hominidae*	Man (only one living species)

IV

MYSTERIOUS CRATERS OF THE CAROLINA COAST

A STUDY IN METHODS OF RESEARCH

By Douglas Johnson

Columbia University

The flat coastal plain of South Carolina and portions of adjacent states is pitted with a vast number of curious oval craters. These depressions vary in longest diameter from a few hundred yards to two or three miles, are partially surrounded by rims of fine-grained sand, and usually have marshes or peat bogs on their floors (Fig. 26). Both sandy rims and marshy floors are commonly covered by pine forest, and the traveler could easily pass close to scores of the craters without recognizing their existence. To the natives the oval craters are known as "bays," a name possibly derived from the bay tree frequently found growing in the depressions.

To the geologist, and particularly to one interested in methods of research, these forms have a double interest. The question of their origin presents a peculiarly intriguing problem, one that has attracted the attention of several investigators with diverse results. And the solution of that problem affords an exceptional opportunity to test the value of different methods of prosecuting research. In the space at my disposal I propose to deal simultaneously with both aspects of the "bays problem," but with primary emphasis upon the question of method.

In developing the discussion it will be most convenient to

ignore the actual sequence of steps followed in studies carried out in the field, and to adopt a logical rather than a chronological order of presenting evidence, arguments, and conclusions. In fact, I can best set before the reader those considerations respecting methods of research which seem to me of paramount importance

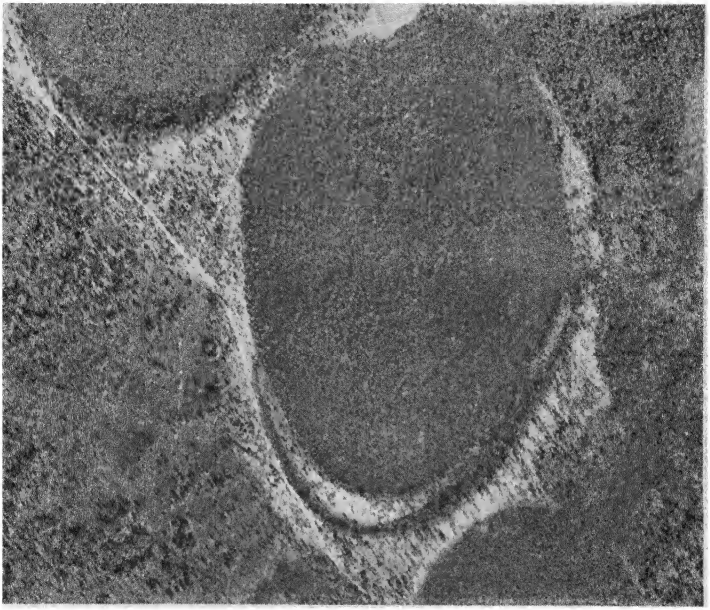


FIG. 26. Oval crater, or "bay," showing sandy rims and inner marsh deposits, both covered with open pine forests. Southeast is toward the bottom of the picture. (Photograph by Fairchild Aerial Surveys, Inc.)

if I completely transform the actual study into an imaginary study. I shall, therefore, take the reader with me on an imaginary excursion into the realms of scientific investigation, promising not to burden him with too many technical terms, but to hold the discussion within bounds of ordinary experience and common-sense reasoning.

One warning, however, is necessary. Since this is to be an imaginary investigation of the Carolina craters, I shall repeatedly represent as my own certain field studies made by other workers, and certain conclusions reached by other students of the problem. In doing this I have a double purpose: first, to simplify the study by concentrating attention on one thing only—methods of research; and second, to be in position to criticize methods whenever desirable without appearing to criticize individuals.

First Step: Observation and Inference

The simplest and most direct approach to the solution of a scientific problem is to observe the facts, and then draw the appropriate inferences from them. Where the facts are clear and their explanation obvious, this method may give trustworthy results. In the case of the Carolina "bays" let us imagine that I observed the following pertinent facts, and drew from them the inferences recorded below:

Each bay or crater is oval (Fig. 27), as if produced by some object striking the earth obliquely and scooping out an elongated depression. The long axes of the oval depressions are almost if not completely parallel, and trend uniformly from northwest to southeast, as though a great shower of large objects coming from the same direction had produced the tens of thousands of craters actually observed in the field and on aerial photographs of the Carolina coastal plain. There is usually a ridge of sand or sandy rim around each crater, commonly most strongly developed about the southeasterly half of the crater as though an object coming from the northwest had pushed most of the excavated debris toward the southeast. Some craters have double (Fig. 26) or even triple rims of sand, as if two or three objects had struck nearly in the same spot, throwing up concentric sandy ridges.

These facts and inferences lead almost inevitably to a single simple interpretation, apparently competent to explain all the facts observed—a great cluster of giant meteorites moving from northwest to southeast and striking the earth at an oblique angle

scooped out the numerous oval craters and piled up debris, chiefly about their southeastern ends. Where two or three meteorites

FIG. 27. Series of oval craters showing parallelism of long axes. The irregular white patches are clearings in the forested country. (Photograph by Fairchild Aerial Surveys, Inc.)

struck close to the same spot double or triple rims of debris were formed.

Fortunately I discover in the field another group of facts which enables me to fix approximately the period at which the

cluster of meteorites struck the earth. Some of the oval craters occur in association with long, nearly straight or but slightly curved ridges of sand of a type found wherever the sea has been receding from a land area but building beaches as it retreats. The characteristics of such beach ridges are so peculiar and so well known to students of shore forms that no competent investigator is likely to confuse them with other phenomena. In the present case I observe that where oval craters are found in association with the beach ridges the craters are often irregularly formed, and the encircling rims poorly developed or wholly lacking. From this I infer that the craters are older than the beach ridges, and that some of the meteorites struck offshore, forming craters on the shallow sea bottom, which were later partly destroyed during the building of beach ridges by wave action.

As many of the craters are very perfect in form and show little evidence of alteration, the fall of the shower of meteorites must have occurred at a comparatively recent period as the geologist reckons time. But since the craters are older than the beach ridges, we must allow sufficient time since the meteoritic shower for waves to build a succession of ridges covering a belt some miles in breadth. I conclude that the shower may have occurred some thousands, possibly some tens of thousands of years ago, but certainly not before the latter part of the glacial period.

Thus I have solved the origin and age of the Carolina craters by a method as simple as it is direct. Were you to ask me to tell just what this method is, I might respond in either of two ways. I might say that it is just the good old-fashioned, common-sense method of observing the pertinent facts and then deciding how they were caused. Were I inclined to be more technical I would say that it is the *inductive method*, according to which the investigator is led directly and almost unconsciously from the observed facts to their more or less obvious explanation. When a problem is simple, the facts clear, and their explanation obvious, the inductive method may give good results.

Second Step: Checking Conclusions

If I am a careful investigator I will not accept my own conclusion without attempting to check its validity. In scientific research one of the commonest methods of making such a check is to determine what consequences, previously unconsidered, ought to follow if the conclusion be correct. Should further study show that these anticipated consequences actually do occur, confidence in the conclusion will be powerfully confirmed.

This so-called *deductive method* of research supplements admirably the inductive method previously described. The conclusion reached by the direct or inductive method is not to be too hastily accepted, but is treated merely as a hypothesis to be tested, a "working hypothesis" to serve as the basis for further study. From this working hypothesis the investigator "deduces" the consequences which logically ought to follow in case the hypothesis be valid, and then ascertains whether the deduced consequences are really matched by facts previously observed or newly discovered as a result of further search. A close correspondence between deduced consequences and observed facts inevitably gives the investigator added confidence in his interpretation.

In my study of the Carolina craters I proceed to apply the deductive method. If it be true that the craters were produced by a great shower of giant meteorites coming from the northwest and striking the Carolina coastal plain at an oblique angle, it should logically follow that the area of abundant craters corresponds with an area of abundant "finds" of meteorites as reported by collectors of these visitors from outer space. Examination of maps showing the distribution of known meteorite finds, published in Nininger's volume, *Our Stone-Pelted Planet*, reveals the fact that the southeastern United States is an area where meteorites have been most abundantly discovered. An oval encompassing the area of most abundant craters also encompasses the area of most abundant meteorites (Fig. 28).

A further logical deduction may be drawn from the hypothesis in question. If meteorites coming from the northwest struck the earth obliquely to form oval craters, they should frequently have penetrated the loose sandy soil of the coastal plain and be now reposing below the surface at or near the southeastern ends of such craters. Since many meteorites contain significant amounts

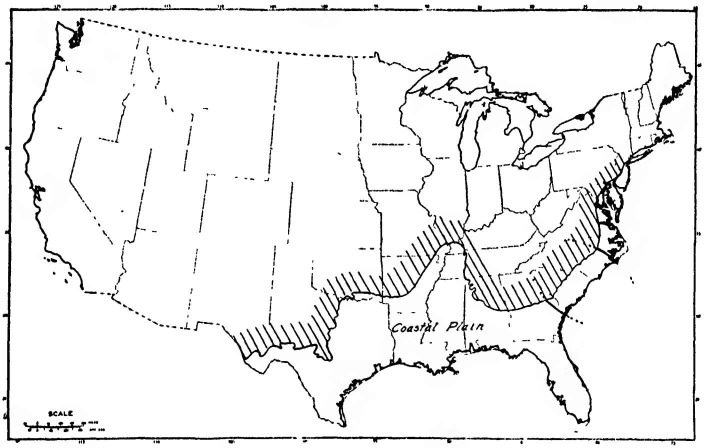


FIG. 28. Outline map showing dotted oval which encloses area of abundant craters and area of most abundant meteorite finds.

of iron, a magnetometer survey of these southeastern areas should frequently show the presence of "magnetic highs," due to the attractive influence of iron within the buried meteorites. Accordingly I make magnetometer surveys covering the southeastern ends of several bays and adjacent territory. In every instance the surveys show distinct magnetic highs in the areas in question.

I conclude that the explanation of crater origin reached by the inductive method, and strikingly confirmed by two critical tests formulated by the deductive method, is well substantiated. On the face of the record I have made out an excellent case for my hypothesis. If the conclusions are published they will be incor-

porated into textbooks of geology and astronomy. But I hesitate. Years of experience in scientific research breed in the investigator an attitude of healthy skepticism. He has learned that only too often "things are not what they seem." He knows that methods of research can be seductive as well as deductive or inductive; that the deductive method is not always as conclusive as it appears to be; that despite widespread belief to the contrary, the supposedly simple and safe inductive method may be quite as dangerous as the deductive method. So the moment he distrusts conclusions reached by the simple and direct inductive method, or by the more indirect deductive method, or by both methods combined, the investigator turns to a third method, far more elaborate than the other two, but far more likely to lead him to the correct solution of his problem.

Third Step: Analysis of Observations and Inferences

This third method of research is best called the *analytical method*. In plain language it consists in separating into its component parts every step in the investigation, and testing critically the validity of each and every part. While one half of the mind, the creative half, searches out and brings together the bricks, the mortar, and the timbers of evidence and argument, and builds these into a complete and satisfying theory, the judicial half of the mind stands jealously by, subjecting to impartial but relentless scrutiny the quality of every brick, the cementing power of every bit of mortar, and the soundness of every timber that goes into the final structure. In the very nature of the case such a process cannot be as direct and rapid as the inductive method by which my meteoritic hypothesis was reached. Nor is it even as simple as the deductive method by which that hypothesis was tested. It is, indeed, far more elaborate than either of these, a complicated, even a cumbersome, method if you please. But it has merits possessed by neither of the other two.

In the first place it makes use of both induction and deduction far more extensively than do the simpler methods which bear

those names. As we shall see, not one interpretation but several are treated both inductively and deductively, thus multiplying whatever advantages may lie in those procedures.

In the second place, the new method brings into play the analytical powers of the mind to an extent impossible in either the inductive or the deductive method. It is this outstanding peculiarity which causes the third method to be known as the "analytical method." And it is the dominating role played by analysis in this method which makes it more likely than any other to discover hidden weaknesses in a hypothesis or theory under investigation.

In the third place, the analytical method involves a conscious effort of the mind to invent and test all possible explanations of a given phenomenon or group of phenomena, thus greatly enhancing the probability that the correct explanation will be discovered. It is this invention and testing of more than one tentative explanation or hypothesis which causes the method to be frequently called the "method of multiple-working hypotheses."

In the fourth place, the analytical method, by its employment of multiple hypotheses, naturally paves the way to discovery that a given phenomenon or group of phenomena is the product of several agencies acting in conjunction or in succession. The true solution of a problem has often remained long hidden because the investigator assumed that the explanation was single and simple, when in reality it was multiple and complex.

Finally, the analytical method possesses the enormous advantage of constantly directing the mind toward new facts and new lines of argument, this being an inevitable and valuable consequence of the effort to test a variety of diverse hypotheses. Under such circumstances the chances that significant facts will escape the eye, or important lines of reasoning escape the mind, are greatly diminished.

Let us now attack the problem of the Carolina craters anew, placing dependence this time on the more complex but more trust-

worthy analytical method. I first go back to my initial step, and subject the original observations to critical analysis.

I stated that the craters are oval, and so they appeared to be. But more critical scrutiny reveals the existence of systematic departures from the simple oval form. The northeastern sides of the bays are prevailingly more strongly curved than the southwestern sides (Fig. 26). It is difficult to see why plunging meteorites should produce craters which are systematically asymmetrical.

Further scrutiny shows that many of the craters are pear-shaped rather than oval, with the narrower end always directed toward the southeast (Fig. 29). It is conceivable that an obliquely plunging meteorite might produce a narrow groove where it first touches the earth, and a broader depression where the full body enters; but in that event the narrow ends of the craters should point toward the northwest.

Parallelism of the northwest-southeast axes of the elongated craters appeared to be one of their most remarkable characteristics. But careful measurements reveal the fact that there are many divergences in axial direction, amounting to more than 50° in some instances, to 80° in a few. It is difficult to understand how meteorites could maintain uniformity of direction for long distances through space to arrive as a cluster, then suddenly to diverge so widely as these angles indicate. Disruptive explosion shortly before striking the earth offers a partial but not wholly satisfying explanation.

It was stated that rims of sand were most strongly developed about the southeastern ends of the craters. Here again my observation was defective. Critical study of more than a hundred craters on the ground and of a far larger number in aerial photographs proves conclusively that the locus of major accumulation is about the *southeast quadrant* of the craters, rather than about the southeastern half. In other words, there is prevailingly a marked asymmetry in the distribution of the sandy rims. It is

FIG. 29. Pear-shaped craters. Irregular white patches are clearings in pine forest. (Photograph by Fairchild Aerial Surveys, Inc.)

difficult to see why a plunging meteorite should push up more debris on the southeastern side of a crater than on the southwestern side.

I pointed out that some craters have double or triple rims. Here observation was incomplete. Further examination reveals

FIG. 30. Aerial mosaic showing very perfect crater interrupting sequence of beach ridges. Southeast is toward the bottom of the picture. (Part of mosaic by Fairchild Aerial Surveys, Inc., for Ocean Forest Company, Myrtle Beach, South Carolina.)

that some craters have as many as six or eight rims. A further significant fact is revealed: while the successive rims are nearly concentric, they converge and ultimately merge before reaching the northwestern ends of the craters (Figs. 26 and 32). About the northwestern ends there are seldom if ever multiple rims,

and frequently no rims at all. It is scarcely conceivable that as many as six or eight successive meteorites should strike almost exactly in the same spot, yet with a systematic slight displacement toward the northwest to give rims so peculiarly arranged as those observed.

I noted that some of the craters associated with beach ridges were irregular in form, and that their rims were incomplete or missing; and it was concluded that the craters must have been formed before the beach ridges were developed. Analysis shows that in this case both observation and reasoning were defective. Beach ridges as well as craters are in places poorly developed or wholly absent. Critical study shows that beach ridges never continue unbroken and well developed through a poorly outlined crater; but that perfectly formed craters repeatedly interrupt the sequence of ridges (Fig. 30). The conclusion seems inescapable that the craters, not the beach ridges, were formed last. Further observation shows that craters are often arranged in chains, like oblong beads strung obliquely on a thread, and conforming perfectly to the trend of the beach ridges. It is difficult to believe that falling meteorites could so perfectly adjust their points of impact to a preëxisting topography.

Thus critical analysis of the very first stage of our investigation places the problem in a wholly new light. It shows how faulty were my initial observations, and hence how untrustworthy were the inferences based on those observations. Grave doubt is thrown on the validity of the conclusion reached by the inductive method of study.

Fourth Step: Analysis of the Deductive Tests

But what about the deductive portion of the study, and the strong support this testing process brought to the meteoritic hypothesis of crater origin? To answer this question the deductive study must now be subjected to critical analysis.

It was pointed out that if the meteoritic hypothesis were correct the area of abundant craters and the area of abundant mete-

orites should correspond; and it was shown that both craters and meteorites are most numerous in an oval area in the southeastern United States. But analysis is inquisitive. It is not satisfied with a vague oval encompassing a vast area. It asks for a more precise map of the distribution of abundant craters, for a similar map showing the distribution of abundant meteorite finds, and for a

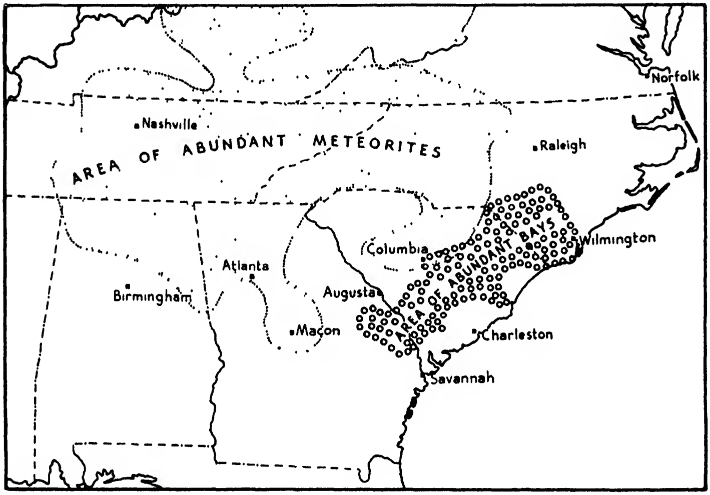


FIG. 31. Map showing area of most abundant meteorite finds, and area of most abundant craters, or "bays."

critical comparison of these two maps. When this is forthcoming (Fig. 31) it is discovered that the area of bays and the area of meteorites are mutually exclusive, save for a very limited overlapping near Columbia, South Carolina. It develops further that in not a single instance has a meteorite ever been found in or near one of the craters, whereas in true meteorite craters such association is common. In the glaring light of analysis one supposedly strong confirmation of the meteoritic hypothesis disappears like the mists of the morning.

Next I turn the light of analysis upon the magnetometer survey which seemingly gave another strong confirmation of the meteoritic interpretation of crater origin. Immediately I note a defect in my method of conducting the surveys. Instead of selecting a typical area containing a number of craters and surveying that entire area, I conducted limited surveys about the southeastern ends of selected craters. This seemed necessary in the interests of economy of time and money. But it jeopardized the intended test of the hypothesis, since if for any reason there were numerous deposits of iron below the surface throughout the region, one would be pretty sure to get one or more magnetic highs somewhere in the vicinity of the southeastern end of almost any crater. So long as my surveys were restricted to the southeastern ends of craters the evidence of the magnetic highs would seem very convincing. Only in case the surveys were extended over a large area, revealing magnetic highs in many other positions, would the inconclusive character of the evidence become apparent.

In the light of this analysis I begin to extend the partial surveys over broader areas. Then I discover that the magnetic highs do occur beyond the areas previously mapped. I find, further, that some of them are beginning to show marked elongation in a northeast-southwest direction. I recall that this is the trend of Appalachian structures in older rocks known to be buried beneath those coastal plain beds in which the craters occur. I know that the coastal plain is a wedge-shaped deposit with its thin edge toward the interior, barely covering the older rocks near the thin inner edge but deeply burying them near the coast line. So my reasoning runs that if the magnetic highs of my surveys are due to iron in the older rocks they should be very strongly developed in surveys made near the inner edge of the coastal plain, but scarcely perceptible in surveys made along the coast. Test surveys about craters in both areas give strong highs in the interior, faint highs near the sea. The necessary conclusion is that the magnetic highs are caused by iron in the older rocks, and not by iron in buried meteorites. What was apparently the strongest support

of the meteoritic hypothesis vanishes under the searching light of analysis.

The analytical method must be applied not only to observations and the inferences drawn from them, and to deductive tests of those inferences, but also to any assumptions that may underlie the inferences. There is one basic assumption underlying the whole discussion of the meteoritic hypothesis which has thus far gone unchallenged. This is the assumption that meteorites striking the earth obliquely will excavate elongated, more or less oval depressions. Let us now analyze this assumption.

Tests made with guns firing large and small projectiles show that a swiftly moving projectile striking the earth obliquely may act in several different ways: (*a*) it may penetrate the ground in the oblique direction of incidence; (*b*) if the angle of incidence is too low, it may strike the surface a glancing blow and pass off into space again; (*c*) it may explode upon impact. In the first instance the crater produced will depart but slightly from the circular form. An angle of impact sufficiently low to produce a greatly elongated crater will not permit penetration. The glancing blow resulting from low-angle impact produces an elongated crater due to the ploughing action of the projectile. In both instances the breadth of the crater is not much greater than the diameter of the projectile. This means that to produce the Carolina craters by simple penetration or by ploughing we must invoke the aid of truly gigantic meteorites having diameters measured in thousands of feet, six to eight thousand or more in some cases. The largest meteorites known to have reached the earth measure less than a score of feet in maximum diameter.

A small projectile may by violent explosion produce a large crater. It is conceivable that a moderately large meteorite by exploding on impact, or by causing either moisture in the ground or the rock material to volatilize with explosive effect, might produce a much larger crater. But explosion craters are circular in outline regardless of the angle at which the projectile strikes the earth. In the World War it was found impossible to deter-

mine from an explosion crater the direction from which the enemy's gun fired a shell. All known meteorite craters are approximately circular. Elongated forms occur only when there is evidence suggesting that two meteorites struck close together.

Thus the advocate of the meteoritic hypothesis finds himself between the two horns of a dilemma. To get elongated craters he must invoke meteorites of utterly unheard-of proportions. To get large craters from smaller meteorites he must invoke explosive impact, which is incompetent to produce elongated craters like those observed.

Fifth Step: Invention and Analysis of Multiple Hypotheses

Analysis has by now abundantly demonstrated that a major error was made in my method of attacking the problem of the Carolina craters. I assumed that the problem was a simple one, easily solved by the simple and direct method of induction, supplemented by the fairly simple but indirect method of deduction. Analysis has convinced me that the problem is far more difficult and complicated than at first supposed, and that its solution can safely be attempted only with the aid of the somewhat cumbersome but far more dependable analytical method, involving the invention of multiple hypotheses and the critical use of analysis during every stage of the procedure.

My next step in prosecuting the study is to analyze further the facts observed, with a view to discovering whether the problem is really a unit, or a series of problems not subject to any one simple explanation. It is first noted that there is the problem of the crater-like depressions, and I deliberately invent as many hypotheses as possible to explain them: (*a*) solution sinkholes resulting from local removal of soluble rock such as limestone, marl, gypsum, phosphate rock, rock salt, arkosic sandstone; (*b*) "blowouts" due to erosive action of the wind; (*c*) local removal of sand by upwelling springs fed by water under artesian pressure; and so on. I also retain (*d*) the hypothesis of excavation

by impact of meteorites, since the fact that this hypothesis was supported by invalid evidence and arguments does not necessarily prove that the hypothesis itself is invalid.

Second I note that the *shape* of the depressions may be a problem entirely distinct from their original creation. Accordingly I invent as many hypotheses as I can to account for their shape, among them the following: (*a*) erosion by waves on lakes occupying the depressions to give their smoothly curving inner margins; (*b*) attack by waves driven by storm winds from the northwest or southeast to explain elongation of the craters in that general direction; (*c*) more wave attack from the southwest than from the northeast to account for the greater bulging and sharper curvature of the northeastern sides of the craters; (*d*) slumping of water-saturated sand into ever-enlarging, artesian-spring basins to give circular depressions which would then be elongated in a northwesterly direction by migration of the springs "headward" or up the coastal plain slope.

In the third place it is noted that the sandy rims may be divided into two groups: (1) those found wholly outside the depressions, often irregular in shape; and (2) more regularly formed ridges which seem to lie within the contours of the craters. As hypotheses of origin for the outer and often irregular rims I entertain (*a*) the possibility that they were thrown out by meteoritic ploughing or explosion; (*b*) that they are ridges of dune sand blown out of the depressions by winds. For the inner and usually very regular rims there is (*a*) the possibility of a wind-blown origin; (*b*) also the possibility that they are beach ridges built by wave action when lakes occupied the craters, their convergence and disappearance toward the northwest being tentatively ascribed to progressive elongation of the lake in that direction due to headward migration of artesian springs or some other cause.

Thus a large number of working hypotheses are quickly accumulated, some of which it is hoped will contain enough of the truth to direct me toward the correct interpretation. To discover these grains of truth, to develop them fully, and to follow them

to a successful solution of the mystery of crater origin is a long and complicated process. Every one of the many working hypotheses must be fully elaborated, its logical consequences deduced, the deduced consequences compared with facts already observed or newly discovered by additional search, invalid hypotheses rejected and faulty hypotheses revised, while at every step of the procedure the searchlight of analysis is constantly directed upon each bit of evidence and each line of argument, in the hope of revealing some weakness hitherto concealed. The hypothesis which survives so drastic a scrutiny does not necessarily represent the correct conclusion. But it is far more likely to be correct than is a conclusion reached by the simple inductive method, the deductive method, or by both combined.

Obviously it would be quite impossible in limited space to present in full all the stages of so complex a process as the analytical study of the Carolina craters. Furthermore, it is not profitable for the investigator to elaborate for others all the excursions, many of them fruitless, he was compelled to make in the course of a given research. I shall, however, present briefly a few of the most interesting deductions developed in the course of my analysis, and the observed facts which seemed to match those deductions so closely as to justify the ultimate conclusion that the mysterious craters of the Carolina coast have, in fact, a complex rather than a simple origin.

Sixth Step: Deductive Testing of Multiple Hypotheses

If the craters are solution sink holes there should be some correlation between distribution of the craters and the distribution of known soluble rock. Furthermore, sink holes of all sizes, down to those measuring but a few feet across, should be associated with the larger craters; and some at least should show the steep, clifflike walls found in many sink holes. Investigation shows that in some localities crater distribution is coincident with the known extension of moderately soluble arkosic sands or more soluble limestone or marl. Elsewhere such correlation is less clear.

Where very soluble rock is present the craters are associated with typical sink holes of all sizes, often in great numbers; and both craters and smaller sinks show steep inner walls, sometimes thirty feet or more in height, a feature which escaped observation until analysis directed search for the pertinent facts. Whatever the cause of the craters, it seems a reasonable inference that the process of crater formation sometimes involved solution of the country rock, with consequent development of typical sink-hole features.

If the craters were excavated in loose sand or loam by springs welling up from below under artesian pressure, loose sand or loam should be found at the surface in areas where the craters are abundant. We should also find that the geological conditions are favorable to the widespread development of artesian water, and it would seem reasonable to expect that some of the artesian springs might still be operating at the present time. Field examination long ago showed that the surface formation of the coastal plain is almost everywhere a loose sand, usually white or buff in color. The bays are most commonly excavated in this sand, or in a sandy loam immediately underlying it. It has also long been known that the Atlantic Coastal Plain offers highly favorable conditions for the development of artesian water supplies, some of the artesian horizons lying close to the surface and giving birth to artesian springs which well up through the overlying sand or sandy loam. Many of these springs are still functioning, and are known to the natives as "fountain springs" or "boiling springs" because of the force with which the water rushes up into the bottoms of the spring basins. Some of the springs are found in major craters, others in minor depressions. All these facts were known to geologists years ago; but their possible relation to the problem of crater origin was not suspected until that problem was recently attacked by the analytical method.

It is only reasonable to suppose that when the coastal plain was first raised above sea level, and before rivers cutting into the plain offered lower outlets for underground waters, there must have been countless such springs bubbling up through the sandy

soil. As the sand must then have been saturated with water it would slump or flow into the ever-enlarging basins, producing circular craters in case the position of the springs remained fixed. But since the coastal plain slopes southeast, and underground waters must then have moved southeastward down the slope, springs would normally migrate "headward" or upcurrent, just as a waterfall in a river migrates upstream. Such migration should transform circular spring basins or craters into more or less oval craters. Where headward migrating springs increased in volume of outflow as solution enlarged their underground channels, excavation should become more effective and the basins should increase in diameter with the passage of time, thus producing pear-shaped craters with the narrower ends directed toward the southeast.

In either case the long axes of the craters should be parallel to the direction of underground-water movement. This would normally be from northwest to southeast; but local conditions might in places deflect the ground-water currents, and hence give craters with long axes diverging notably from the usual downslope direction. Thus past geological conditions, not considered until the analytical method was applied to the problem of crater origin, appear to explain at one stroke the major forms of the craters, the prevailing northwest-southeast orientation of their axes, and the occasional marked deviations from that prevailing direction.

Upwelling waters tend to seek the lowest outlets on the surface. In a series of beach ridges and intervening swales such outlets will normally be in the lower swales. But sometimes these swales are occupied by impervious silts and clays. Under such conditions the easiest route to the surface would be through the porous sands of the broad, relatively flat ridges. In these facts we apparently find explanation of the existence of chains of craters parallel to the trend of the beach-ridge system, the chains some times occupying the swales, sometimes occurring on the ridges.

Another legitimate deduction from the artesian spring hy-

pothesis of crater excavation deserves attention. If the craters were excavated by artesian waters welling up from under ground, those waters must have had a means of escape on the surface. In other words, there should have been outlet channels leading from the craters, and those channels should still be visible where spring waters are still flowing, or where they ceased to flow so recently that the channels have not yet been buried by drifting sand or other deposits. Craters produced by scooping or by explosion, and surrounded by rims of the expelled debris, normally do not have outlet channels. So long as the meteoritic hypothesis dominated attention no one sought for outlet channels, and apparently no one noticed their existence. Not until the analytical method was applied to the problem, and one of its multiple hypotheses by requiring their existence directed observation to their discovery, was it learned that outlet channels are one of the characteristic features of the craters (Figs. 26, 27, and 32.) Some of the channels are still occupied by outflowing spring waters.

The outflowing of water from the craters offers a reasonable explanation for a fact seemingly inexplicable on the basis of the meteoritic hypothesis. Calculations of the amount of sand found in the ridges or rims surrounding the craters show that the total is insignificant compared to the vast bulk of material removed to form the basins of the craters. This fact, seemingly fatal to the meteoritic hypothesis, is readily explained if during the whole period of crater development outflowing waters were busy transporting seaward the silt and fine sand composing part of the original coastal plain deposit.

The hypothesis that lakes occupied the craters opens the way to explain certain of their features quite independently of their original formation. One characteristic of the depressions is the remarkable simplicity or smoothness of their inner contours. Slumping of saturated sand might give roughly circular or ovoid basins, but it is difficult to believe that this process alone would produce such smooth outlines as are actually observed. It is known, however, that waves tend to simplify shorelines by

eroding projecting angles and filling reëntrants. If the basins were occupied by lakes it seems inevitable that their inner con-

FIG. 32. Large pear-shaped crater showing eight or more parallel ridges or rims, and outlet channel. A sand dune, partially covered with vegetation, has encroached upon the western portion of the crater. (Photograph by Fairchild Aerial Surveys, Inc.)

tours must have been regularized by wave erosion and wave deposition.

But while waves tend to smooth out the irregularities of a

shore, dominant wave action from one direction may erode one side of a lake more than another. It will later appear that in the region of the Carolina bays dominant winds come from the northwest, west, and southwest. Winds from the first direction would tend to drive waves against the southeastern end of the lake; but winds from the west and southwest would both drive waves against the northeastern side of the lake, the maximum effect being felt toward the central portion of that side because there the waves would arrive from the greatest fetch of open and deeper water. The result should be to sharpen the curvature of that side of the crater. We have already noted that such asymmetry of crater outline does exist.

The lake hypothesis makes possible a full explanation of the parallel sandy ridges within the craters. These have all the characteristics of ordinary beach ridges cast up by waves, a fact not noticed by any observer until the analytical method with its systematic use of multiple hypotheses was brought into play. Such ridges are characteristically parallel or nearly so for limited distances, but normally converge toward their extremities. Migration of the lakes toward the northwest, due to spring migration "upcurrent" or up the ground-water slope, would prevent the development of successive beach ridges about the northwestern ends of the craters, where they are in fact normally absent. Gradual lowering of the ground-water level, an inevitable result of the progressive deepening of river valleys in the coastal plain, must have caused a lowering of lake levels. This fully accounts for the fact that we sometimes find as many as six or eight successive and nearly concentric beach ridges, the inner ones progressively lower than the outer ones (Fig. 32). Where soluble rocks are involved the ground-water level, and hence the level of near-by lakes, may drop suddenly when new channel ways are opened by solution. This offers a possible explanation of the fact, hitherto overlooked, that in certain closely associated groups of craters there uniformly occurs a break between an upper and a lower beach ridge or set of beach ridges.

Just as the analytical study of the Carolina craters first opened our eyes to the possibility that the craters and their sandy rims might have quite independent modes of origin, so it was the analytical method that first directed attention to the possibility that the sandy rims themselves might belong to two distinct categories of phenomena each having its own peculiar mode of origin. Once the observer had his "out-sight" sharpened by this fuller insight into the nature of his problem, he had no difficulty in distinguishing two types of ridges or rims: one type prevailingly narrow, seldom over two or three feet high, regular in outline, and located within the crater margin; and a second type often of great breadth, from three to fifteen feet or more in height, with exterior borders highly irregular and poorly defined, the whole lying on or outside of the crater rim (Figs. 26, 29, 32). As a rule the contrast is not obvious, but critical study seems to establish the reality of the distinction.

It manifestly is not possible to ascribe to wave action those ridges or rims lying outside the crater and sometimes extending one or two thousand feet from what must have been the border of the lake occupying the crater basin. The sand is fairly fine, of uniform grade, and the surface of the deposit often has the undulatory aspect of an accumulation laid down by the wind. Thus the hypothesis that the sandy ridges are wind-blown deposits seems applicable to the outer rims. It may explain in part the inner rims, where these appear to be dune ridges superposed on earlier beach ridges in a manner common along sandy shores. But the aeolian hypothesis (wind) alone fails to explain both the form and the distribution of the inner ridges, while the lacustrine hypothesis (lake) satisfactorily accounts for both. Only when applied to the outer rims does the aeolian hypothesis appear competent. An adequate source of bare, dry sand was provided by the sandy beaches bordering the lakes in the craters.

Analysis suggests two deductive tests which we may apply to the aeolian hypothesis:

(*a*) If the outer rims consist of material brought by the wind,

the deposit should contain no material too coarse to be transported by that agency, except such as Indians or other forms of life might have imported. Under the meteoritic hypothesis there should be fragments of underlying rock layers ploughed up by the plunging meteorite or blown up by explosion. Careful study of many rims by different observers has failed to reveal the pres-

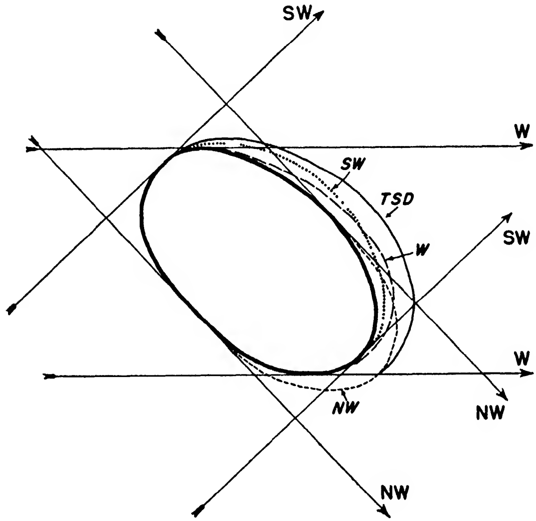


FIG. 33. Diagram illustrating ideal oval crater showing wind-deposited sand rims which should result from dominant northwest winds (NW), west winds (W), and southwest winds (SW). The master rim resulting from the total sand deposit is marked TSD.

ence of any rock fragments in the rim, except occasional bits of chert or other similar material believed to have been carried in by Indians. The rims are composed of fine sand identical in character with known dune sands.

(*b*) It is the outer rims which show best the major accumula-tion of sand about the southeastern quadrants of the craters, a peculiar asymmetry difficult to explain under the meteoritic hy-pothesis of crater origin. If the outer rims are really of aeolian

origin, examination of weather records for the Carolina coastal plain, or for that part of the plain occupied by the craters, should show dominant winds from the opposite quarter. Study of records furnished by the U.S. Weather Bureau shows that throughout the region in question winds of fairly high velocity blow for the greatest number of hours from northwest, west, and southwest. If we plot about an ideal oval bay (Fig. 33) the rims which would be formed by winds blowing from each of the three directions mentioned, and then combine these into a single master rim (TSD, Fig. 33), we find that the sand accumulation has its major development about the southeast quadrant of the ideal crater, precisely as it does in nature about the southeast quadrants of the Carolina craters. The hypothesis that the large outer rims are of aeolian origin is thus strongly confirmed.

Final Step: Conclusion

In our present discussion it has been shown that the inductive study of an apparently easy and simple problem led us to a solution dramatic in its quality and convincing in its simplicity. Deductive tests of that solution appeared strongly to support its validity. But critical analysis stepped in and revealed hidden weaknesses. Application of the full analytical method, with its invention and testing of multiple hypotheses, led us to conclude that the problem was difficult rather than easy, complex rather than simple, and that a combination of unrelated causes was necessary to account for the facts observed. The meteoritic hypothesis was found wholly incompetent to explain the Carolina craters and their associated features. Evidence was discovered which seemed to indicate that the craters resulted in part from the solution of soluble rock formations and in part from removal of sand by artesian springs; that wave action on lakes within the craters reshaped their shores and built some of their sandy rims; and that wind action built up the major rims partially surrounding the outer margins of the craters. For the simple meteoritic hy-

pothesis it became necessary to substitute the more complex "solution-artesian-lacustrine-aeolian hypothesis."

Whether or not this more complex hypothesis is valid it is too early to say. The case in its favor is stronger than I have represented, since there are lines of supporting evidence which for lack of space cannot be set forth. But the hypothesis has its weaknesses. Headward migration of artesian springs has been assumed, not demonstrated. Excavation of large basins by artesian springs has not been observed, or at least not recognized as such. To some the perfection of form of many bays may appear too remarkable to be fully explained by the processes invoked. I can only say that the explanation offered is the best I have been able to devise, and that to me the forces appealed to seem competent to produce the results observed. But I fully realize that some other hypothesis, perhaps one that wholly escaped my search, may prove the key which will solve the mystery of the Carolina craters.

Be that as it may, some worth-while lessons can be drawn from our study. One is that the simplicity of an explanation is no guarantee of its validity. The human mind prefers simple explanations of natural phenomena. Yet it remains true that Nature often moves in complex as well as in mysterious ways her wonders to perform.

We have seen that the inductive method of research, often championed as safer than the deductive method, can lead us quite as far astray. It is as easy to induce erroneous inferences from observed facts as it is to deduce erroneous consequences from invented hypotheses.

We have demonstrated the value of properly controlled deductive reasoning. Uncritical deductions lent a false color of validity to an erroneous hypothesis. But deductions controlled and tested by critical analysis revealed our mistakes, and paved the way to a fuller understanding of the problem.

We have gained an impression of the dominant role which analysis should play in scientific research. The successful solution

of any complex problem will usually depend primarily upon the skillful employment of the analytical powers of the investigator's mind.

Finally, we have had to confess that not even the analytical method can guarantee the investigator success in his search for truth. Had we examined other methods of research, including the experimental method so valuable in the physical sciences, our conclusion must have been the same. There is no royal road to truth, nor does that goddess bear any patent of nobility by which she may be recognized when found. Only too often is the truth of today revealed in tomorrow's light as error in disguise.

V

HOW THE EARTH SHOWS ITS AGE

By Alfred C. Lane

Professor of Geology, Emeritus, Tufts College

In considering how the earth shows its age, emphasis should be placed on the word "show." As the age of a horse is shown by the patterns and the wearing down of the teeth, we find similar significance in the "dents" and "sierras" of the earth (Fig. 34). Young mountains, like those in California which are still growing and experiencing shocks known as earthquakes, are sharp, while older mountains of the Canadian shield, for example, those around Great Bear Lake, are worn down and rounded (Figs. 35, 36).

The teeth patterns of the successive types of horses in the evolutionary series become increasingly complex, just as the patterns of a mature horse are more complex than those of a young horse. This is even more true of the successive types of elephants. H. F. Osborn and E. H. Colbert have suggested that one might almost make a scale of geologic time with the aid of the complex lines made by the dentine on the huge molars of an elephant. If straightened out, these lines would be found to have increased in length in accordance with the stage of evolution of the elephant and the wearing down of the tooth. Thus a scale of time for that part of the Tertiary period during which they existed might be constructed. (Fig. 36.)

One of the common ways of determining the age of trees is, of course, by counting the annual rings. These show some of the big California trees to be more than three thousand years old,

Bands that resemble these tree rings have been found in beds of rock and some of them are considered to be annual deposits; they

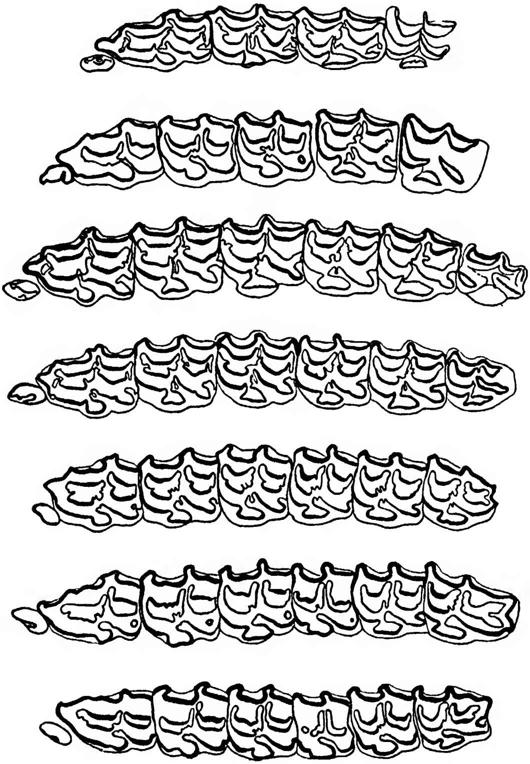


FIG. 34. Illustrating dentition patterns of fossil horses. *Plesippus shoshonensis* Gidley. Enamel patterns of left upper cheek teeth in seven selected individuals arranged according to age. In the first two dentitions, the premolars belong to the deciduous set. Hagerman lake beds, upper Pliocene, Idaho. (After Gazin.)

are called "varves," by geologists as suggested by the Swedish geologist DeGeer (Fig. 37).

There are three methods of measuring time which can also be used in geology—by progressive, by periodic, and by paroxysmal

activities. Time measurements have been summed up by a poet in the following lines:

> "And still the burning sands within the glass do fall.
> And still the hands around the dial creep
> And still the water clock doth drip and weep
> And that is all."

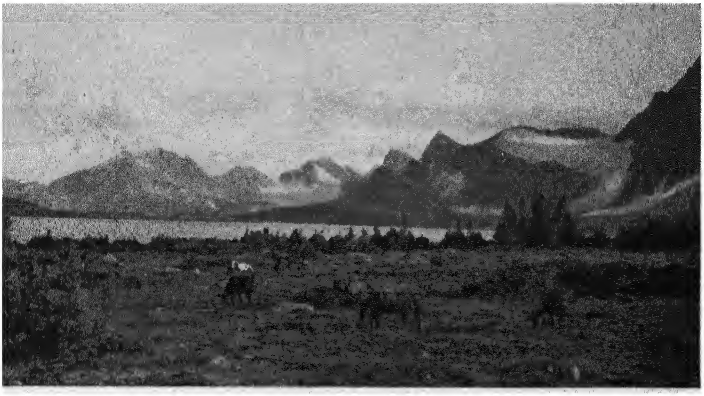


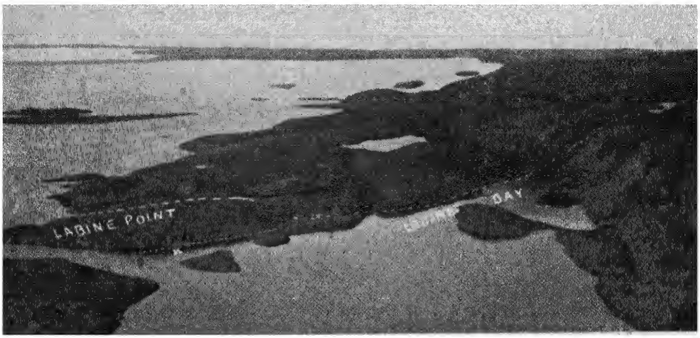
FIG. 35. Mountains of Longuin Valley, Jasper National Park. (Courtesy Canadian National Railways.)

There is progressive action in the falling of the sands in the hourglass, and this is paralleled in the tilting or eroding of land, which may eventually make a mighty canyon of a small gully. These are motions with no reversal (Figs. 38, 39). Omar Khayyám described their irrevocability when he said:

> "The moving finger writes and having writ
> Moves on, nor all your piety nor wit
> Can lure it back to cancel half a line
> Nor all your tears wash out a word of it."

Periodic action is seen in the swing of the pendulum causing the movement of hands around a clock, or the alternation of day and night, new and full moon, summer and winter.

Finally, there is action by regular paroxysms in the striking of a clock, or, as in the days of the Romans, in the continual dripping of water from the water clock. Progressive action is involved, too, when the drop-forming, surface tension of the water is overcome at regular intervals by the accumulating strain of gravity which pulls the water down. The eruptions of Old Faithful geyser are famous among regular paroxysms, but there are many other examples, seen in the earthquake, avalanche, or landslide, where



Photograph by Royal Canadian Air Force.

FIG. 36. Mountains of the Canadian shield around Great Bear Lake. (After Spence.)

gradually accumulating stress repeatedly reaches a breaking point.

To get any definite idea of the age of the earth, beyond a general impression that it is very old, three things are necessary, which are found even in a three-thousand-year-old Egyptian slate ruler: a definite starting point, a unit of measurement, and a means of counting these units.

Geologists consider the beginning to be the present, and the present to be the key to the past. The beginning from which measurements are made should be sharply defined, unique, and easy to recognize. All these adjectives apply to the present, except for the fact that the date is slowly shifting. However, the

shift is so slight in relation to the long periods of time which are to be considered that it is insignificant.

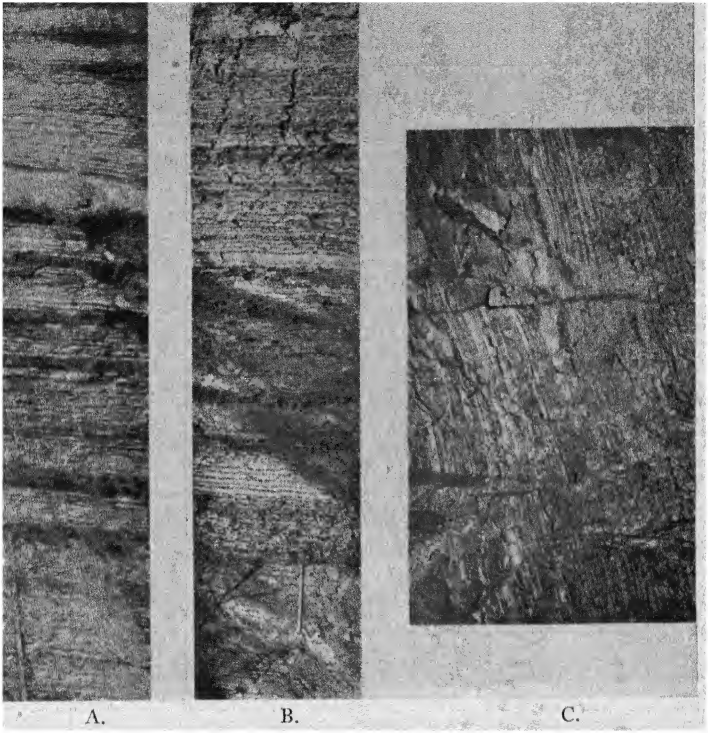


FIG. 37. Yearly rings, or varves, due to annual deposits of sediment. Sunspot banding is superimposed on the varves. A. Alternations of fine-grained gray shale with limestone layers. B. Alternations of fine-grained clay shale and layers with calcareous nodules. C. Argillaceous limestone. All from Upper Devonian. See Fig. 44. (After Korn.)

If a beginning point is sought which is more sharply defined geologically than the date of the birth of Christ, which, indeed, was determined hundreds of years after his death, it may be found in the spread of man over the world since the last ice age, which

has produced profound differences in the distribution of plants and animals throughout the world, and caused wide scattering of associated objects.

A more sharply defined starting point than the first appearance of an animal species is the time when the species first becomes numerous. The author clearly remembers having seen a windrow

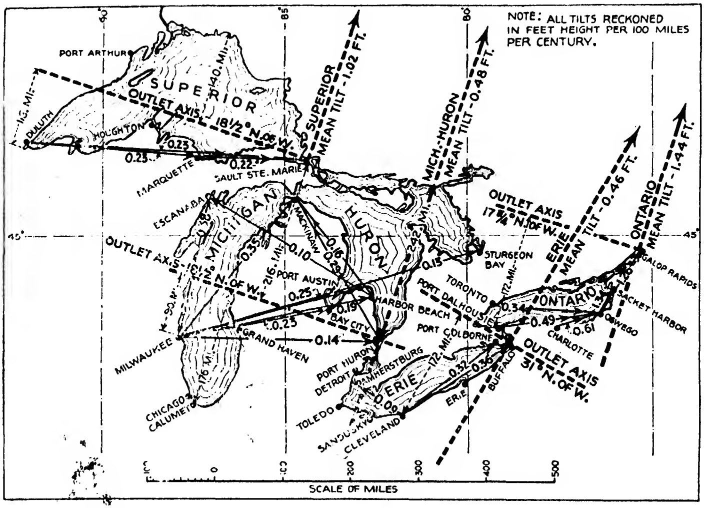


FIG. 38. Rates and directions of earth tilt for each of the Great Lakes. Deduced from comparisons of change in heights of water during the past fifty years, indicated by pairs of gages at localities shown on map. Maintained by the U. S. Lake Survey and others.

of Colorado beetles, two or three inches high, on the Atlantic shore about 1880. This beetle took its name from the fact that in the year Colorado was admitted to statehood, these insects swept eastward across the United States from the habitat where they had been peacefully munching wild potato leaves since time immemorial. A million years hence the stratum marked by the beetle wings may possibly have become a sharp geological datum.

The year as marked by the rings in dicotyledon stems, fish

scales, and other growth bands, cannot be our only unit of time. A unit marked in the rock strata, in which we are striving to recognize the longer periods, must also be sought (Fig. 37). One of these longer periods is the period of the sunspot, which can be traced back to the days of Joseph and his fat and lean kine.

The double sunspot cycle is a period upon which C. G. Abbott of the Smithsonian Institute has been doing detailed work. Sun-

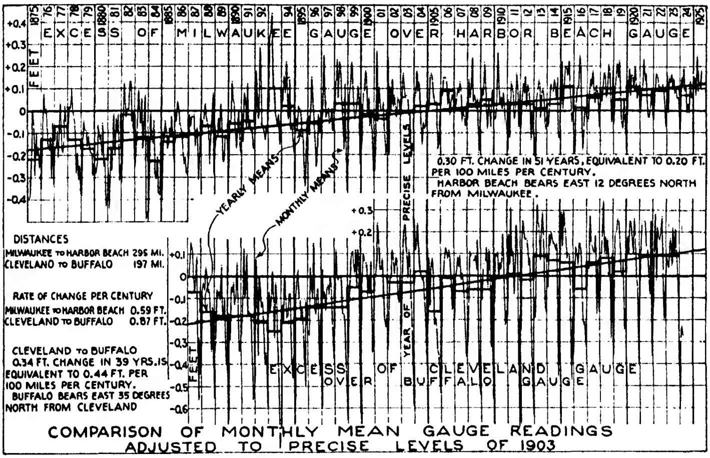


FIG. 39. Diagram presenting two separate determinations of the rate of tilting. (After Freeman.)

spots appear farthest from the sun's equator, in pairs—one positive and one negative in electric charge, and one of the pair nearer the equator. As they move toward the equator, they grow larger and more abundant and then gradually diminish. When the eleven-year cycle is nearly finished, other spots will appear, but their electrical signs are now the opposite of what they were in the preceding eleven-year cycle. Thus a complete electrical cycle takes between twenty-two and twenty-three years, though the abundance of the sunspots varies in a cycle of about eleven years. They seem to be due largely to the pull of Jupiter, which takes

about the same length of time (11.9 years) as that of the sunspot cycle to travel around the sun, though of course other planets have an influence, and the length of the cycles is not precisely the same (Figs. 40, 41).

Sunspots have been used to prophesy weather and other changes, from the time of Sir Norman Lockyer to that of H. Janvrin Brown and H. H. Clayton. It is not easy to say what the climatic effects of such astronomic cycles will be, because, as Clayton and others have pointed out, while climate in a particular spot does vary with the sunspot cycle or other astronomic cycles, the centers of the continents behave differently from the margins. A place along the coast, which for a time may have an increase in temperature and rainfall as the sunspots increase, may suddenly have a decrease as the centers of action shift. There is a complex of secondary effects, as the direct effects express themselves in changes in wind and ocean circulation. Moreover, changes that affect the amount of ice on land and in icecaps also affect the sea level and consequent oceanic circulation (Figs. 42, 43).

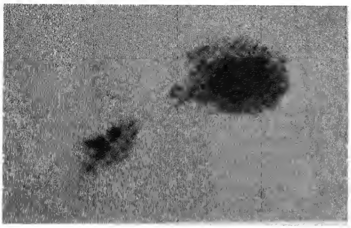


FIG. 40. Bipolar group of sunspots as described on page 113. (From *The Telescope*, 1936.)

These sunspot cycles can be traced in the thickness of tree rings and in varve bands of sediment, not only in present growths and recent beds but also in those of the Tertiary, as described by W. H. Bradley, and in the Devonian, as found by H. Korn. Insofar as the rhythm of the eleven-year cycle can be found superimposed on the varves, the varves can be interpreted more certainly as yearly layers (Figs. 37 and 44).

But still longer cycles are desired. The one which has attracted most interest and is most promising is the precessional cycle, which depends upon the fact that the axis on which the earth

spins is not at right angles to the plane of its orbit. At present the South Pole is inclined toward the sun when the earth passes nearest to it in its elliptical course. As a result the southern hemisphere has a hotter summer than the northern. Every 13,000 years or so this trend is reversed, the North Pole inclines toward the sun, and the northern summer is the hotter. The entire cycle takes 26,000 years. This climatic effect is modified by the difference in distribution of land and sea in the two hemispheres. Yet the cycle may well alter the climate of the whole world. Moreover, the earth's orbit is sometimes more nearly circular and at other times more elliptical, varying, according to the pull of the other planets, in a cycle longer and more complicated than the precessional cycle.

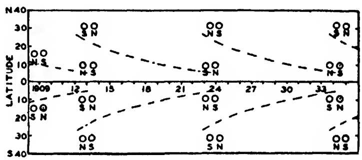


FIG. 41. Law of sunspot polarity. The curves represent the approximate variation in mean latitude and the corresponding magnetic polarities of sunspots observed at Mount Wilson from June, 1908, until January, 1935. (From *The Telescope*, 1936.)

We cannot at present use the precessional or any of the longer cycles as units of geological time. It is a goal of the future to find by induction a long-time rhythm beat superimposed on the yearly rhythms, and then, by means of the yearly rhythms and the thicknesses deposited in a year, to learn the duration of the larger rhythms recognized by regular repetitions in the bands of sediments, some of which have been known as cyclothems and megacyclothems (Figs. 37 and 44).

In fine-grained oil shales, Bradley has found yearly layers that are from .1 to .01 of a millimeter thick. Collet found bands averaging 2.5 millimeters in deposits of Lake Geneva; and Korn, in his recent monograph, records a similar thickness of 2.33 millimeters for the Thuringian roofing slates. Annual layers increase from such figures to several centimeters or more, and the coarseness of the deposits varies with the size of the spring floods which formed them and with their nearness to shore or delta.

In views of strata which one often sees, however, it is clear that if measured in inches or feet the rhythms of fine-grained rocks which resemble annual rings of trees—for example, the flint bands which are a foot to a foot and a half apart in chalk—rarely if ever represent years but probably represent the precessional or some longer cycle.

Another type of cycle is the orogenic cycle, which is supposed to

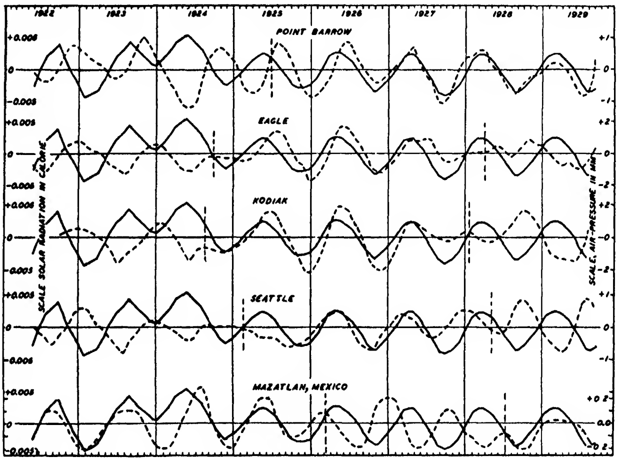


FIG. 42. Solar radiation, eleven-month period (unbroken line). Air pressure (broken line). To illustrate change from positive to negative correlation of sun radiation with weather. (Barometric pressures for various points.) (After Clayton.)

depend upon an accumulation of strain in the earth's crust, relieved at regular intervals, when it passes a certain point, by mountain-building paroxysms with volcanic and earthquake disturbances. Two contradictory causes of such cycles have been proposed. The older theory, which H. Stille continues to advocate, holds that the strain on the earth's crust is produced by a cooling and shrinking of the earth. The other, suggested by Joly and elaborated by Holmes and Jeffreys, and more recently by Kirsch, attributes the energy to the accumulation of heat gener-

ated by atomic disintegration. The author has called it a teakettle theory because the lid of the teakettle may be regularly lifted to discharge steam. Grabau and various authors preceding him give such varying estimates for these cycles that they cannot be used as units. Since the rate of loss of heat under the ocean is not known, we cannot tell whether the interior of the earth is heating up or cooling down. The origin of the earth is still undetermined, and we do not know whether it is hot enough at the center to be

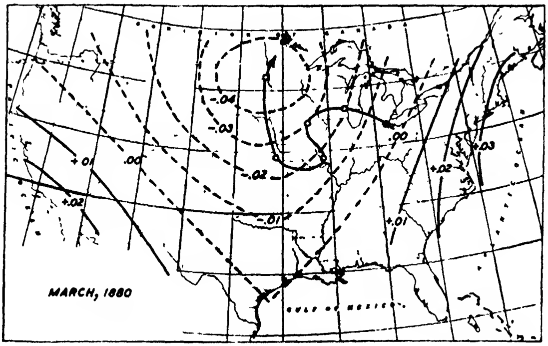


FIG. 43. Map indicating centers of greatest minus departure in period of slightly over two years, showing movement of the centers of oscillation. (After Clayton.)

losing temperature and shrinking, in spite of the heating effect of atomic disintegration.

We must go back to the year as our unit of time.

A standard illustration of a progressive activity used to measure time is the recession of Niagara Falls, which has changed markedly in the last fifty years. The Falls are so obviously cutting back the gorge that it is not surprising that this was recognized very early and that estimates were made as to the time required to form the gorge. The necessary factor is the approximate rate at which the Falls are receding, which has been determined by careful surveys to be about five feet a year.

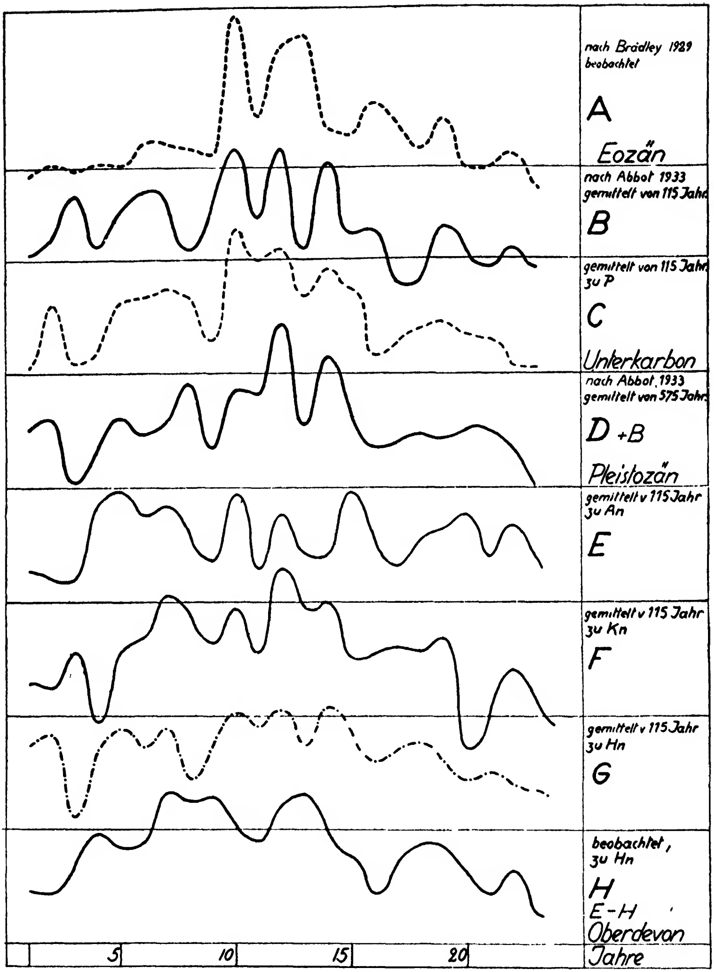


FIG. 44. Results of annual layer, or varve, measurements for the sunspot cycles of 23 years. Arranged by Korn from various authors. See Fig. 37.

It must be remembered that even this recession is somewhat paroxysmal as well as progressive, due to a series of collapses: a

piece gives way now here, now there. A table rock present when the author first saw the Falls in the '80's has since collapsed, and part of the front of the American Falls has disappeared. Now that so much water is being taken away underground for electric power, the rate of recession must be diminishing, but there are quite accurate surveys in addition to earlier vague sketches and descriptions dating back to about 1842 which indicate the average rate. To recede the full length of seven miles at five feet a year would take something like seven thousand years. But a glance at the map (Fig. 45) shows that the width of the gorge is not uniform and that it would be very unsafe to assume a uniform rate of retreat. Through the Whirlpool Rapids north of the railroad bridge the gorge is narrower and shallower, and the large volume of water has difficulty passing through the smaller channel. An explanation of this is derived from an interesting discovery by Grove K. Gilbert, the first to do accurate work. He found that the entire Great Lakes basin is being tilted and is rising to the north, so that when the lake level is compared with points on the rim, the shore points on the northern side are higher, or the lake level relatively lower, and the streams cascade into the lake. Early voyageurs passing the mouth of one such stream appropriately named it Rapid River.

On the other hand, the lake water rose in the valleys of the streams at the southwest end, to be backed up and shut off into separate lakes by beaches formed of sand working along the lake shore. These facts were recognized by Winchell years ago. The Chicago outlet is at the extreme south end and will tend to drain all the water that way. The opening of that outlet caused John R. Freeman, commissioner for the Chicago Sanitary District, to make a careful study of the amount of tilt, which is about half a foot per 100 miles per 100 years (Figs. 38, 39).

In 1926 and 1928 F. B. Taylor reviewed for the Michigan Academy the evidence on the rate of tilting of the Great Lakes basin. This tilting is due to the earth's resilience—its readjusting after the removal of the load of the great ice sheet. The under-

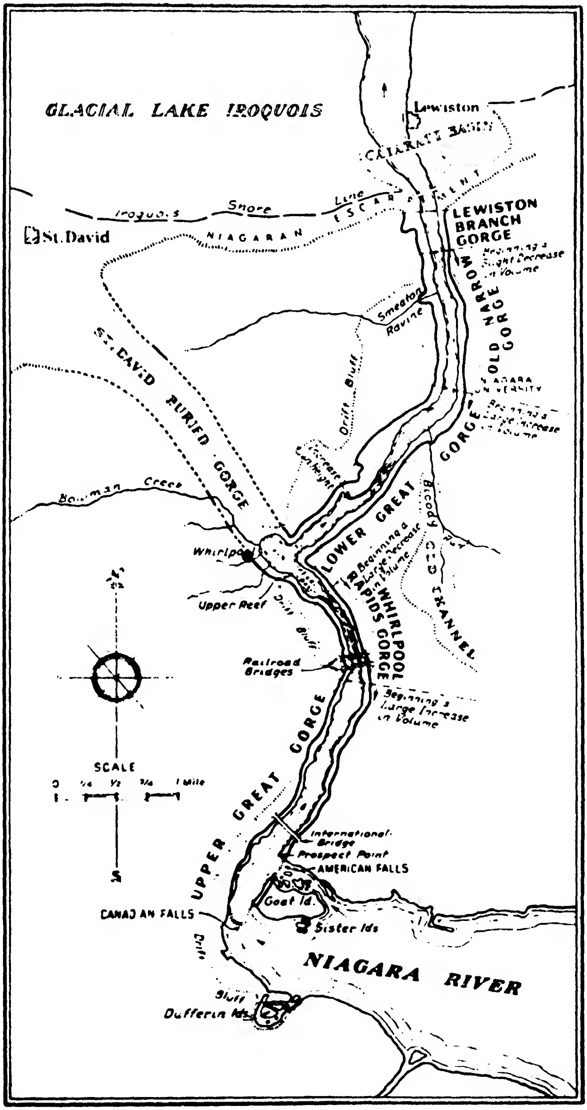


FIG. 45. Diagram of Niagara River Gorge showing relationships to the lakes. (Modified from Gilbert, U. S. Geological Survey.)

lying rock was promptly relieved of the elastic compression due to the weight of the ice, but the downward bending of the surface crust into a weaker, more plastic layer has been relieved but slowly, and this recovery is still going on. Recent studies in Fennoscandia indicate an uplift of some 600 meters already, and from the deficiency in gravity over the area about 50 meters further uprise in the future is indicated. Taylor called attention to the fact that the rate of tilting is not necessarily uniform between the various stations that the engineers have used around the Great Lakes, but is greater to the north. That is altogether probable. He assumed that the tilt is confined to the area north of a hinge line and is uniform there; while south of the hinge line there is no tilt. Actually the hinge line has shifted from time to time, and the tilt may not be uniform north of it. However, Taylor's assumption is probably nearer the truth than the assumption of a uniform rate for the whole basin.

Taylor found a rate of tilt of .43 foot per 100 miles per century from measurements on the beaches of North Bay, Ontario; and the findings of S. Moore, as well as G. K. Gilbert's measurements between Milwaukee, Wisconsin, and Escanaba, Michigan, confirm this. If, however, one supposes that all this tilt is concentrated north of a hinge line passing 84 miles south of Escanaba, the rate would be .98. Freeman's findings of rates in the Lake Erie basin, varying between one half and one foot, are in agreement with the above figures. He estimated the rate for a 172-mile line across Lake Erie to be 0.46 foot per hundred miles per century. If one again supposes that all of this tilt is concentrated along the 33 miles northeast of where Taylor has drawn the hinge line, the rate would be 2.38. Buffalo would then have an uplift of .79. It is plain that the lake level for most of the Lake Erie basin is rising, and E. L. Moseley's studies around Sandusky corroborate this.

Three thousand years ago the land around North Bay in the Lake Huron basin, 238 miles northeast of Taylor's hinge line through Port Huron, must have been more than 100 feet lower

than at present (based on a rate of tilt of 1.58). The upper three of the Great Lakes would then have drained by way of Lake Nipissing (Fig. 46) and the Ottawa River instead of Niagara. This is known to have been the case, and the question has been studied in considerable detail by Leverett, Stanley, and especially by Taylor who estimates that the tilting has been going on for

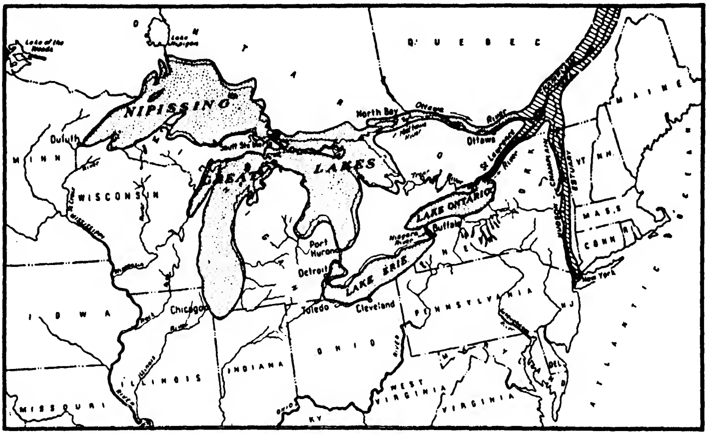


FIG 46. Nipissing Great Lakes, Lake Erie, and Lake Ontario. The North Bay, or Ottawa River, outlet of the Nipissing Lakes; the St. Lawrence River outlet of Ontario-Erie to the Champlain Sea. (After Leverett and Taylor.)

three thousand years. The former continuous body of water of the upper lakes is known as Glacial Lake Nipissing. During its existence much less water went through Niagara—only that from Lake Erie. It is probably not a mere coincidence that the falls were at that time cutting the stretch of narrow gorge known as the Whirlpool Rapids Gorge. We find that further back in geologic time (Fig. 47) there was an ice dam that blocked the Nipissing channel and made a larger lake known as Lake Algonquin. This might account for the wider lower valley. There are further complications, such as earlier glacial lakes like Lake

Warren and Agassiz; but these cannot be considered here.[1]

It is to be noted that the lower part of the valley is wider and has more gentle slopes. Over forty years ago, the Reverend George Frederick Wright of Oberlin made some observations and estimates on the rates of recession and wearing back of the side slopes of the cuts of the Gorge Railway, which he also applied to

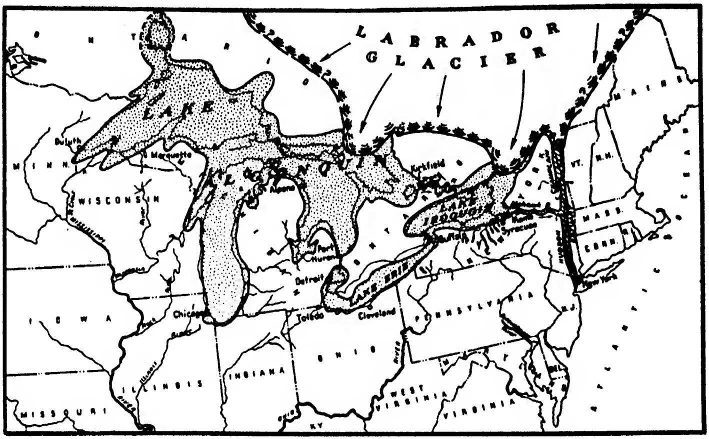


FIG. 47. Glacial Lakes, Algonquin and Iroquois; Lake Erie. Trent River outlet of Algonquin; Mohawk River outlet of Iroquois-Erie to the Hudson-Champlain estuary of the Atlantic Ocean. (After Leverett and Taylor.)

the widening of the whole valley. Evidence provided by the widening of the valley, recession of the falls, tilt of the land, and shift of outlets all agrees as to the order of time involved. Likewise, the drying up of the large lakes in the Far West, described by Jones and to be considered again later; the decay of radioactivity in the radium-bearing tufas of the hot springs of Yellowstone National Park, described by Schlundt; the tempera-

1. The story of the Great Lakes has recently been recorded in some detail and in popular style by Helen Martin of the Michigan Geological Survey in a book called *Ne-saw-je-won*.

ture in the Michigan copper mines, and numerous other features all over the globe agree as to the order of time since the Glacial or Pluvial period.

Erosion and deposition are perpetually going on. A cloudburst in an arroyo can move huge boulders and carve out gullies. As Tennyson says, the forces of nature forever "draw down

FIG. 48. Erosion and deposition as shown in an aerial view of mouth of Fork Canyon, Utah, after floods of 1930. One flood in 1923 and four floods in 1930, which originated on barren spots at the head of the watershed, swept out of this canyon, causing heavy damage to farm lands, highways, and railroads. Geological evidence based upon depth of cutting in Bonneville deltas and deposits shows that these floods are unprecedented during post-Bonneville history. (U. S. Forest Service 264093. Hawkins Photo Co., 1931.)

Eonian hills, and sow the dust of continents to be" (Fig. 48).

One of the longest periods estimated by studying erosion was the time required for the erosion of Bryce Canyon in Utah. Professor Pack of the State University found that some of the trees on the canyon's edge had roots which turned back like a fish hook. From the way they grew he calculated that the canyon wall had

receded at the rate of one foot in fifty years. Since the distance across the canyon was around 100 miles, he estimated that about 26,000,000 years had elapsed since erosion began in the Miocene period.

Besides various estimates of time provided by the action of rivers in working down, backwards, and sideways, there are also

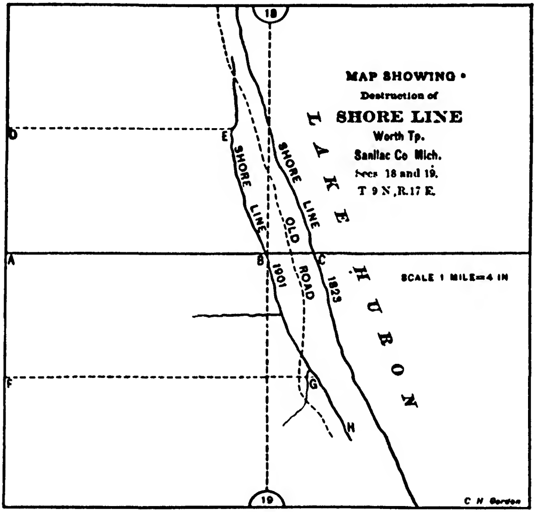


FIG. 49. Map to illustrate recession of Lake Huron shore. (After Mich. Geol. Survey.)

those provided by the action of the ocean in cutting its cliffs, sea arches, and curves, and leaving distinct lines—such as the barnacle lines of Minot's Ledge Lighthouse, which are being observed to learn if there is a slow subsidence. In hard rock there is little erosion, and Minot's Ledge Lighthouse, built in 1858, seems to be quite well protected from erosion by the life which has settled upon it. Softer beds, however, are more rapidly eroded and cliffs of unconsolidated, even glacial, deposits recede fairly fast. The author was much interested to find on the shore of Lake Huron a place where a road, located on old maps of a hundred years ago,

ran off a bluff and reappeared half a mile farther on. There were some plow marks, too: one set which ran off the bluff, another set which ran along the edge but with some room for the off-horse, and finally a third set which again ran perpendicular to the bluff but stopped in time for the team to turn. Twenty acres on which the estate of a Cleveland banker was still paying taxes had gone into Lake Huron (Fig. 49).

ACCUMULATION

As the continents are worn down, the product must be deposited somewhere: there is a deposition credit to match the erosion debit (Fig. 48). The material is carried off partly in solution. River waters contain various salts; if they empty into a lake or ocean, part is deposited. There are coral reefs and marl sludge at lake bottoms, but the more soluble salts may remain and accumulate. Thus lakes without an outlet on or below the surface become salt. The most famous, perhaps, are the Dead Sea, into which the Jordan flows, and Great Salt Lake. That this lake is a shrunken remnant of a greater body of water known as Lake Bonneville is clearly shown by the old shore lines that surround it (Fig. 48).

Professor Jones of Reno, Nevada, made some interesting studies of a similar body of water, known as Lake Lahontan, which at one time occupied a good part of Nevada and now is shrunk to Pyramid and Winnemucca lakes. He tried, in three ways, to estimate the time that has elapsed since it was a great fresh-water lake with an outlet to the north.

First, there were available two analyses of the lake water made about thirty years apart, the later one showing it to be more salty. Second, he made observations on evaporation and found out how long a time would be required at present rates for the evaporation of all the water of the once great Lake Lahontan. Third, from the amount of sodium being brought down by the present streams into Pyramid Lake, he estimated how long the accumulation of the salts would take. All three methods agree in suggesting that

only a few thousand years ago there was vastly more water in Nevada than at present.

The same method on a large scale has been applied by Joly to the ocean of the world. The ocean is a single great salt lake into which the streams are bringing various salts—more than are subtracted by the salt breezes. Joly therefore tried to calculate how long it would take for the rivers to bring in all the sodium that is now in the ocean. The estimate is a rough one, for he lists about eight factors that would make the time he reckoned too great, and an equal number that would make it too small. Since then a new factor has been introduced by the discovery of the capacity of clay to absorb sodium when it is deposited—a problem which A. C. Spencer and K. Murata have recently investigated. Moreover, very little is known about the size of the continents in times past; it may be that during long periods they were very small, or were low and mantled with material already leached, so that little dissolved matter was carried to the ocean. In addition, we have until recently had but little idea of the true contribution of the streams, since few analyses were available that showed what they were carrying in time of flood.

Finally, the questions of where the ocean came from in the first place and how much salt it may have contained originally add an uncertainty. However, we may in fairness say this: first, Joly's argument shows that the ocean is many million years old, even if conditions were always what they are at present; and secondly, some limit to its age is suggested. Moreover, it seems likely that there would have been a gradual change in the composition of the ocean: the more soluble salts like sodium would tend to accumulate while those that are more readily precipitated, like lime, though they might increase, would not increase in proportion. Analyses of the waters found in drilling for oil seem to indicate that there is some truth in this assumption. As a general rule, the older the beds the less the proportion of sodium—unless the beds were deposited in fresh water on the land, or have been leached by water descending from some ancient land surface.

In addition to dissolved material a stream, of course, bears solid material. For every gully eroded there will be an alluvial fan (Fig. 48). The material eroded during a cloudburst is deposited on the plain. The products of river erosion may make a delta, or at any rate will be deposited finally in the ocean. The rate of deposition therefore affords some idea of time.

The most direct method of measuring deposition is that of the archeologists. In an oriental city materials are brought in but nothing is carried out. If the sun-dried clay bricks of a house are ruined by fire, the site is flattened and the next house built on the ruins. At one time children were buried underneath the hearthstone. Thus the town grows upward. El Tainar (Tel-al-Judaidah), not far from Damascus, is a good illustration of the careful work of modern archeologists of the Chicago Oriental Society (Fig. 50). The excavators take off layer after layer, noting carefully the materials and signs of past life that they find. At El Tainar the upper layers contained crosses of the Christian Byzantine period; below that were Roman lamps; below that, coins of Alexander; and further down were signs of the culture of the Philistines and the period of David and Uriah. Successively deeper layers revealed relics of the time of Joshua; relics of the time when Ur of the Chaldees was a great city, and the Magi began those records of which Alfred Noyes sings in the "Watchers of the Sky":

> "The records grow unceasingly, and each new grain of truth
> Is packed, like radium, with whole worlds of light.
> The eclipses timed in Babylon help us now
> To clock that gradual quickening of the moon,
> Ten seconds in a century. Who that wrote
> On those clay tablets could foresee
> His gift to future ages . . ."

In like manner, many Troys have been found on the site of Troy, and Pumpelly found in Turkestan a section in two mounds or kurgans which dated back beyond the iron and bronze age to the

stone age and the time when animals were first domesticated. The rate of deposition was 2 to 2½ feet a century.

Using this scale, we can estimate the end of the last ice age. In its general order this estimate checks with those made from Niagara and the Western lakes in indicating that some 20,000 years ago there was a glacial period in which ice covered much of the earth. In the nonglaciated regions this period has been called the wet, or Pluvial period. The history of civilized man is the story of his spread over the earth since the glacial period.

It must not be forgotten that besides the deposits made by man and by water, there is also deposition by wind, which may help to bury the works of man. The central part of the continents is generally dry, and around the edge of the central dust bowl deposits of dune sand or finer material known as loess are found, which are likely to bury some of the relics of animal or human life, especially those of temporary inhabitants around the edges of the desert such as Pumpelly has studied. At the same time, in other spots the wind may be eroding and taking away the soil. The author has a picture taken by Professor Berkey in the Great Interior Gobi of Asia, showing a wind erosion of about 5 feet in 150 years.

Just as the relics of man in the Tainar mound show evolutionary progress, so it has been with life as a whole. There are cycles of uplift and depression and, accordingly, recurrences of beds of sandstone and finer shales and limestone. But if a study of the life in them is made, it is found that each bed of a given time has its characteristic fossil bones and that in life there has been a steady progress: there has been no recurrence, no sign that an animal once extinct has ever reappeared.

Geologists are able to match the beds in one place with those of another by comparing their fossils, and have gradually built up a succession of beds approximately 400,000 feet thick. Of course they were not laid down in one place; nor are they known from study of any one place, as our deepest borings penetrate only to a depth of 15,004 feet. Even the time required for the deposition

PERIOD I
600-300 A D

The level of an early Christian church, with its chapel and close Byzantine coins and bronze crosses of the priests

PERIOD II
300 A D -64 B C

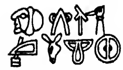
A village partly contemporary with St Paul and early Christian missionary activity in Antioch. Coins of the Caesars and Roman lamps

PERIOD III
64 B C -500 B C.

An occupation of the period of the Persian Empire, showing also traces of the Hellenization of the Orient under Alexander the Great

PERIOD IV
500 B C -1000 B C

Layers of the Syrian Hittite kingdom, contemporary with the later Assyrian Empire and the Babylonian Nebuchadnezzar. Hittite hieroglyphs



PERIOD V
1000 B C.-1200 B C.

Ceramic traces of the "Peoples of the Sea," some of whom are known as the Philistines, others as the Achaeans who sacked Troy

PERIOD VI
1200 B C.-1600 B C

A period of ethnic movements and extensive pottery importation Infiltration of *Habiru* into Palestine and Joshua's capture of Jericho

PERIOD VII
1600 B C -1800 B.C

Evidence of cultural relations with the east, attested by cylinder seals of the Hurrian peoples of northern Mesopotamia, identified as the Horites of the Bible.

PERIOD VIII
1800 B.C.-2000 B.C.

Small painted bowls related to the pottery of the Hyksos or "Shepherd Kings" of Egypt. Time of the Patriarchs.

PERIOD IX
2000 B C -2400 B C.

The beginning of a series of clay figurines of the "Mother Goddess" type, which are remarkable for their intentional grotesqueness.

PERIOD X
2400 B C -2600 B.C.

An occupation by the makers of goblets with fork-scratched decoration an evidence of trade with northern Mesopotamia and central Syria

PERIOD XI
2600 B C -3100 B.C.

Importation of cylinder seals from Abraham's city, Ur of the Chaldees, and fine red and black polished pottery from Asia Minor and the Balkans

PERIOD XII
3100 B.C -3400 B C

Earliest general use of metal, rapid mastery of the technique of casting figures in copper. Decline of flint and bone tools

PERIOD XIII
3400 B.C.-3800 B.C.

Painted hand-made pottery as fine as any subsequent painted style. Flint and bone implements, with earliest traces of copper.

PERIOD XIV
3800 B C.-4500 B.C.

The earliest Syrian village life, with the same material culture as found in near-by caves. Hand-made pottery, bone and flint tools, no traces of metal.

VIRGIN SOIL: SIX FEET UNDER THE PRESENT WATER LEVEL.

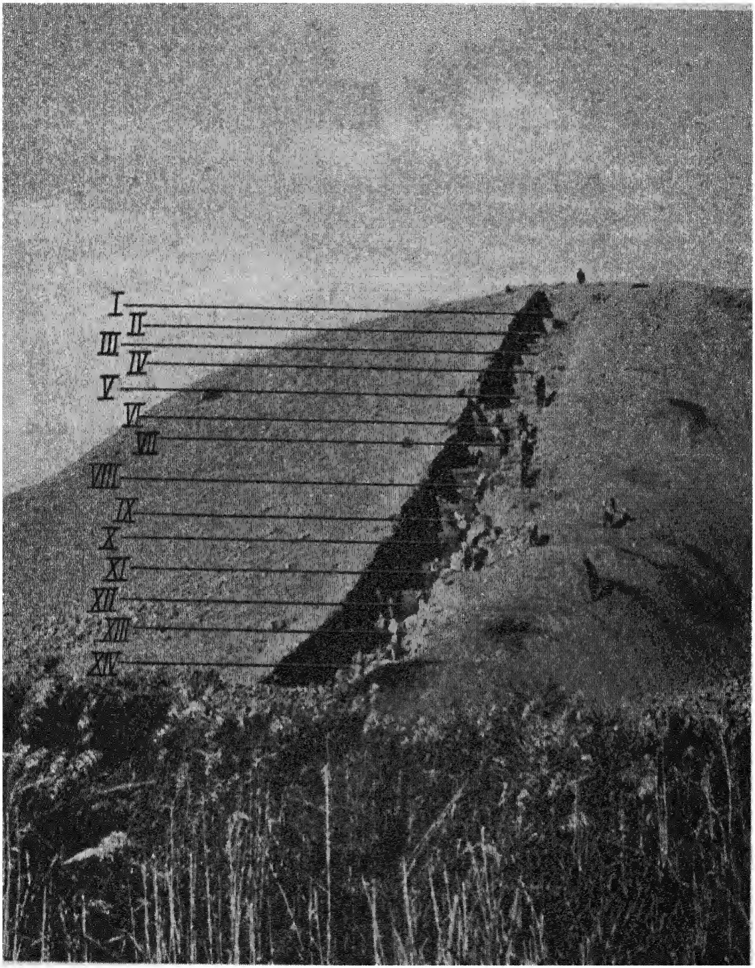


FIG. 50. A "step trench" up the slope of Tel-al-Judaidah, opened by the Syrian expedition of the University of Chicago. Such mounds are entirely artificial, this one resulting from the accumulations of five thousand years of occupation. Each period (indicated by the Roman numerals) contains typical objects, illustrated and described on opposite page, which do not regularly occur in any other period. By a careful check of the objects from a given "floor," the archaeologist can establish their sequence. The step trench shown above established a complete chronology of North Syria for the first time.

of individual beds varies greatly, and the definite figures now available come from what may be called the "popcorn methods" which may now be indicated.

POPCORN METHODS

All the above methods of determining earth history, however, either give us but little of the last chapter of the geologic story, or are very inexact. Generally they do not indicate the age of the earth. They show it must be old, but except for the method based on the salt content of the ocean, they do not challenge the epigram of an early geologist that, in the story of geology, there is no sign of a beginning, no prospect of an end. As often occurs in science, however, study of the *very brief* here advances our knowledge of the *very long*. Similarly, in astronomy the study of the very small has yielded new ideas of the size of the universe, and an improvement in the grain of photographic films assists the astronomer almost as much as would a more powerful telescope.

Suppose you have a wrist watch with luminous figures and also a small pocket lens such as one might use to examine minerals or flowers. Sit in the dark for about two minutes and then focus the lens on the luminous dial. The chances are that you will find that the light is not quiet but flickering. Perhaps you will be able to discern that it is a shower of sparks like a bursting rocket. What causes the light? The explosion of an atom. As the atom explodes, it throws off a particle known as an alpha ray which soon picks up an electron and becomes an atom of helium. This strikes the zinc sulphide and makes it glow. Thus every flash means that an atom has exploded. In such explosions, three kinds of rays may be emitted: alpha particles, which by picking up electrons become helium atoms; beta particles, which are electrons (units of electricity); and gamma rays, which are practically X rays. Most attention will be given to the alpha particles, but similar statements may often be made of the beta (Fig. 51).

It is possible to tell how long a popcorn machine has been running by noting the number of kernels popping per minute,

how much corn is popped, and how much unpopped. The number of kernels popping will depend upon the original amount of corn and the time that has elapsed since it began popping. This is also true of the exploding atoms.

While your wrist watch may show you, by the light the atoms make when they hit the zinc sulphide, that they are exploding, they are too fast to count; but by means of a special instrument

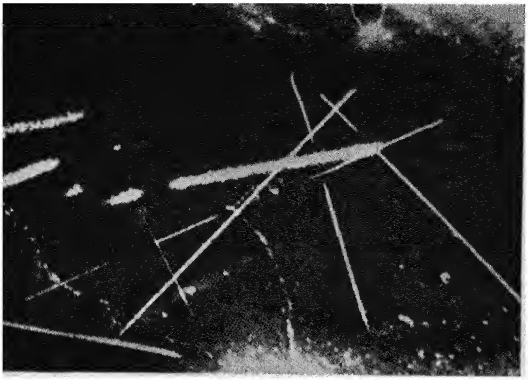


FIG. 51. Reproduction of typical cloud-chamber photograph, showing tracks of various types and ages. (After Rayton and Wilkins.)

known as a spinthariscope or by a Geiger counter they can be counted. A milligram of uranium will yield about ninety-eight sparks a second. Twenty-six milligrams of allanite, a mineral which has only about one per cent of uranium or its equivalent, give twelve flashes every minute. This procedure (bombardment of zinc sulphide), however, is not the way in which the tracks of individual helium atoms are recognized and photographed. The atoms can be recognized and photographed when they are made to shoot through air supersaturated with water vapor. They then cause the water to concentrate in a little line of drops of moisture, which become visible if a side light falls on them while the background is dark (Fig. 51).

Knowing the number of atoms in a gram of uranium I—6.02 x 10^{23} divided by the weight of the uranium I atom compared with hydrogen (for uranium I amounting to 2,500 billion billion atoms to the gram)—and having counted the number of helium atoms given off during a definite length of time, we can easily estimate how long it would take to generate a certain amount of helium and to generate a certain amount of the lead by the disintegration of uranium I. In the process of change from uranium I to lead, not one but eight helium atoms are given off; in the change from actino-uranium seven, and from thorium six. About 152 million millionths of a gram of uranium I changes to lead each year.

Of particular interest to geologists is the fact that regardless of variations in temperature, pressure, chemical combination, or any other variable factor in the earth's outer crust, the explosion of these atoms apparently continues, irregularly in detail, yet on the average uniformly.

When a kernel of popcorn explodes, it gives off a gas. So does an atom of uranium. The corn kernel might conceivably make a little dent in the receptacle. In any event, it produces some energy and so does the uranium atom. Finally, the popped kernel remains. In the case of uranium the "kernel" in itself is not stable but keeps on popping until, finally, the result is lead. The process has several consequences: first, adjacent minerals may be affected, perhaps discolored; second, the parent element changes and disappears; third, new elements are produced. Madame Curie's discovery of the "disintegrating" elements like uranium has been the basis for a whole new science of radiology, which has proved useful not only in medicine but also in geology.

A number of disintegrating or radioactive elements exist in the earth's crust. Their change is so slow that it was overlooked until Madame Curie's time. Actino-uranium is inseparable from uranium. Potassium is the commonest of all. Then there are thorium; rubidium, a rare relative of lithium and potassium; and samarium, which is even more rare.

Zonal discoloration is one of the ways in which the minerals that contain radioactive elements can be recognized. In the rock around them there often appears a puckering, to which B. K. Emerson called the author's attention long before it was known to be associated with radioactivity. The rocks have cracks which radiate from the minerals. Often the radioactive mineral and the rock around it look weathered. This is not unnatural, for uranium has a stronger affinity for oxygen than has the lead that is produced.

The helium bullets have a direct effect, but they do not go very far—a few centimeters in air, only a few hundredths of a millimeter in rocks. They do most of their work at the end of their flight, just as a bullet will make only a small hole through glass if it is discharged near by, and perhaps shatter the pane near the end of its course. Minute rings, known as halos, are found under the microscope around specks of radioactive material. Each radioactive element gives off its helium bullets with a different velocity. Thus they have different ranges. Wilkins has measured their tracks and ranges (Fig. 51).

Roughly, the more rapidly the elements explode, the greater the velocity and the greater the range of the ejected helium bullet. Thus a series of rings will be developed which correspond to different radioactive elements (Fig. 52). G. H. Henderson has studied these. Uranium produces a whole series of elements, each of which, theoretically, would have its own particular ring; moreover, uranium is not really one element but two, which differ slightly in atomic weight. Uranium I weighs 238, and actinouranium weighs 235. Nier has recently determined the exact proportions of the two. U^{238} is 139 times more abundant than U^{235} and is the parent of the famous radium, whereas U^{235} is the parent of another series, the actinium series, and disintegrates so much more rapidly that today there is a much smaller proportion present than there was in the geologic past. Half of what exists today will disintegrate in 713,000,000 years, whereas uranium 238 will decrease by half only in 4,560,000,000 years. Therefore,

in the older minerals and rocks there should have been a greater proportion of the actinium series, and the halos produced by that series should be well developed. G. H. Henderson has estimated geologic time by measuring the intensities of the rings, with a rough correspondence to more accurate methods (Fig. 52, 1, 2).

Finally, the well-known fact which started radium investigation may be mentioned: the minerals containing radioactive elements affect a photographic plate even in the dark, and, figuratively expressed, take their own photographs. A flattened surface,

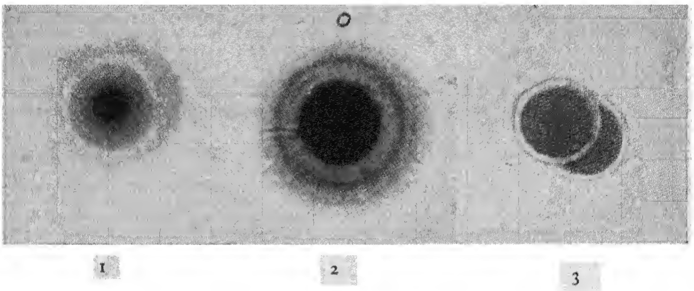


FIG. 52. 1, 2. Photomicrographs of thorium halos. (After Henderson.) 3. Photomicrographs of overlapping uranium halos. (After Kerr-Lawson.)

not necessarily polished, if left in the dark long enough (it may be a day or a month) will reveal its nature; and a polished specimen of radium ore from, say, a vein in the very old, worn-down mountains of Great Bear Lake, Canada, gives a beautiful photographic pattern (Fig. 53). As Muench recently stated, the main difficulty is to keep the plate exposed to a weak mineral long enough without its being disturbed. A strong ore of uranium like the Great Bear Lake pitchblende takes only a day; but some of the weaker minerals containing only 1 or 2 per cent of uranium or thorium require a month or more, and it is sometimes difficult to find a place that will not be disturbed in that time by someone who wants to clean up or dust!

The second effect produced by a radioactive element, which

has already been mentioned, is the actual decay of the original element. Generally we do not know how much of any such element existed in the beginning, but it is possible to determine how long ago U^{235} and U^{238} should have existed in equal quantities, and this may have been at the earth's beginning. It can also be

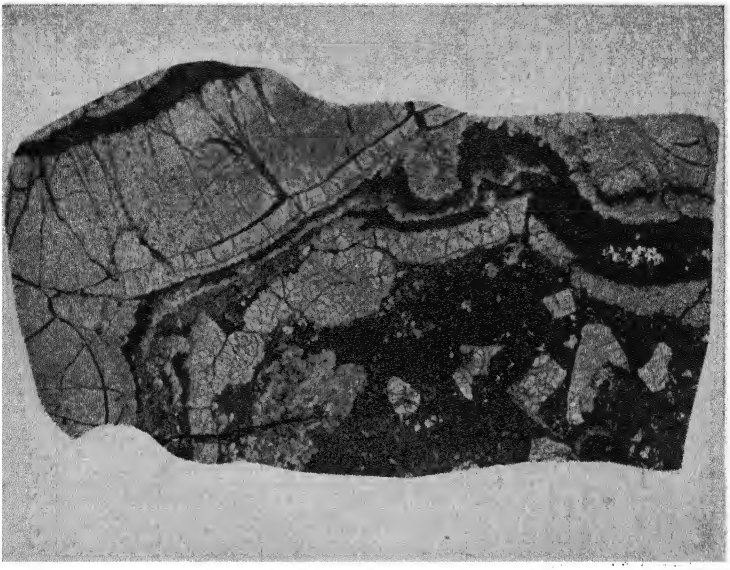


FIG. 53. Radiogram of pitchblende ore from Great Bear Lake, about 2½ inches long. The dark part is largely quartz, with some of the finer speckles consisting of galenaor, chalcopyrite, etc. The dark part is the effect upon the photographic film of bombardment from the pitchblende; exposure 48 hours or more. (After Spence.)

said that, judging by the proportion of lead produced from uranium to uranium today, the earth must be less than ten billion years old. If it were older there would be more of this lead. The more accurate our knowledge becomes concerning the amount of lead, uranium, and thorium yet remaining in the crust of the earth, the more definite is the upper limit that can be placed on the age of the earth. Estimates have been made by Henry Norris

Russell and H. Jeffreys, A. Holmes, and R. D. Evans, but it is difficult to determine accurately the average proportions of radioactive elements in the earth.

Working with radium, which is much shorter lived and much more active than the uranium from which it is formed, H. Schlundt has successfully estimated ages by its decay. He started with the fact that the various mineral waters, for instance those of the Yellowstone National Park, are radioactive because they contain radium and not uranium. The deposits (tufas) made by the springs in the park are also radioactive. Now, assuming that the radioactivity of the water has been constant, and that of all the deposits when first laid down was the same, then 1,680 years later the radioactivity would be half as strong, and 3,360 years later a quarter as strong. It is therefore possible to determine the age of a tufa by comparing its percentage of radioactivity with that of the tufas now being deposited.

Schlundt tested various tufas and calculated that some of the older ones, which have been overlaid by a moraine formed when ice filled the valley, might be about 14,000 years old.

The third consequence of radioactive disintegration, much more important for making accurate reckonings, is the production of new elements and minerals. Uranium minerals, especially uraninite, which is the richest ore of uranium and is black, frequently are surrounded by a zone of brown, red, orange, and yellow minerals. This is well illustrated at Katanga, in the Belgian Congo, and at the Ruggles mine in New Hampshire which is being studied by Shaub. Most important is the change of uranium to lead. Initially a pure uranium salt contains only uranium, but soon other elements build up. Except for the final products of disintegration, the proportions of the elements formed from the uranium become fixed. In the same way, when water flows through a series of tanks, the amount in any tank will depend upon the rates at which the water runs in and out, and if the flow continues long enough a position of equilibrium will be reached in which the level will be fairly constant. In the uranium minerals

equilibrium condition is attained in about a million years. The age of very young minerals may be determined by the degree to which this equilibrium has been attained, that is, by the per cent of equilibrium, the ratio of radioactivity or radium to the parent uranium or thorium. This ratio has been determined, however, in but very few cases. One is that of some yellow scales occurring in a Wyoming gypsum, a mineral known as dakeite, which turned out to be only a few hundred years old.

After equilibrium has been established, however, the proportion of the intermediate elements to the parent is independent of age, and the key to the age is the ratio of the lead to any one of them. If then the amounts of both the parent and the lead can be determined, and the lead distinguished from any lead that is not radiogenic and not derived from the parent, the age can be figured with some degree of accuracy. The first estimates of age were based directly on uranium and lead analyses. Later work has proved that these estimates were pretty accurate for a number of minerals which did not contain much thorium or lead derived from anything but uranium. But soon it was found that thorium was sometimes an important contributor of lead. The problem also arose as to how to distinguish radiogenic lead from other lead that might have found its way into the mineral or rock. The careful laboratory work of Richards and Baxter proved that the atomic weight of radiogenic lead was quite different from that of ordinary lead of galena veins, which is always about 207.21. Only within the last year has that been found to vary slightly, as it naturally should, since one can hardly expect to find it absolutely free from radio lead. Baxter, Richards, and their colleagues promoted the testing of these leads and found some with an atomic weight close to 206, but only in minerals or rocks, like Swedish kolm which is practically free from thorium. Thus some much more reliable estimates of age were obtained. But the effect of the actinium series was still an uncertain factor. For some time it has been known or suspected that the actinium series started from an isotope of uranium, that is, atoms which are included in

uranium but either differ from those which produce radium or, owing to some quirk, undergo a different transformation.

This matter has been settled only in the past year or so, by the aid of wave lengths on the spectrograph. This instrument has undergone considerable development since the days when Bunsen recognized the lines of potassium and sodium, and the whole spectrum was only a matter of a few inches. The length of the spectrum has been increased to meters, spectrum lines from invisible rays and X rays have been recorded, and computations made of the dependence of the line upon the atomic weight of the atom that makes it. Modern spectra include an enormous quantity; hundreds of thousands, of lines arranged in series and bands. Frequently bands from different compounds or elements may fall on top of each other so that their unraveling is a job for experts, who sometimes find it difficult to distinguish them.

Dr. John L. Rose probably deserves the credit for first separating the lines of the three kinds of lead that are produced by atomic disintegration—radiogenic leads with atomic weights, 206, 207, and 208, respectively. However, he was not able to separate a line 204 from the line 207. It was possible for him to show that the spectra of radio leads are markedly different; that the radiogenic lead from Katanga is overwhelmingly 206 lead, but that it contains some 207 lead; and that, as von Grosse had suggested, it should be possible to estimate the age of the mineral by the ratio of the 207 lead, which is derived from actinium and known as actinium D, to the radium lead 206 (Fig. 54).

This had already been a subject of calculation, and the author had suggested that there was a fivefold, or perhaps even a sevenfold, check on the possibility of estimating the age of some large uraninite crystals from Wilberforce, Canada. But there were fundamental uncertainties, since it was not known exactly what atoms are present in uranium nor their proportion.

The mass spectrograph of Nier is a highly developed machine compared with the early spectroscopes of the days of Bunsen.

Briefly, it is an apparatus in which positive ions, i. e., atomic groups, are formed by the collision of a beam of electrons (particles of electricity) with the vapor or gas from the substance under investigation. If necessary (when a relatively nonvolatile substance is used), it can be heated in a small furnace inside the tube of the spectrometer. UCl_4, UBr_4, and PbI are the substances used in investigations concerned with the ages of minerals. The uranium or lead ions, carried along by the negative elec-

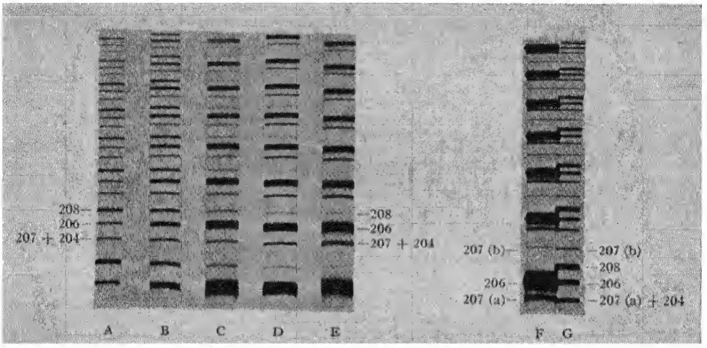


FIG. 54. Geologic time and isotopic constitution of radiogenic lead. A, ordinary lead; B, Colorado carnotite; C, Wilberforce uraninite; D, Great Bear Lake pitchblende; E, Katanga pitchblende; F, Katanga pitchblende; G, ordinary lead. (After Rose and Stranathan.)

trons, are speeded up as they pass from one slit to another and then sent around a magnetic analyzer, that is, put between two poles of a magnet which tend to deflect them from their course. The heavier they are, the less they are deflected. Thus, as Rose has shown, they would strike a photographic plate at different points. Nier has them enter an electrometer tube amplifier which can be moved along, and which sums up the amount of energy of the ions passing through a slit. The results are photographically recorded or summed up in any way that the physicist finds convenient. Thus he gets a curve of observations which shows, for

any particular kind of atom, the amount of energy and therefore the abundance of the different atoms in the various points of the spectrum.

Although actino-uranium is only $1/140$th part of the element uranium, it is responsible for over $1/25$th of the radioactivity

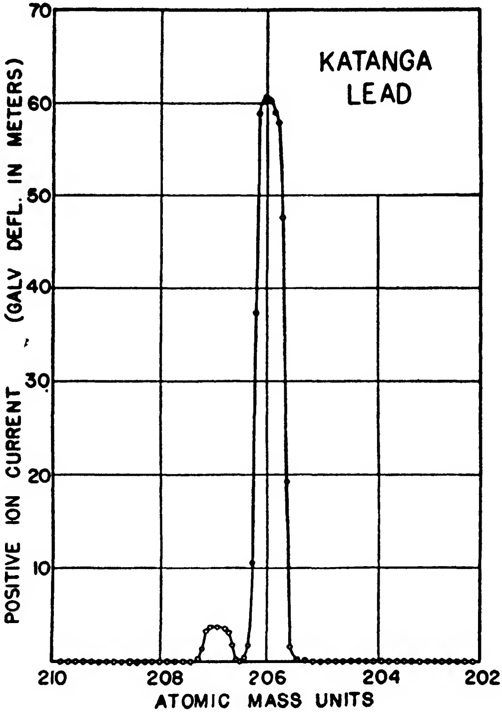


FIG. 55. Mass spectrum showing isotopes of uranium lead. The experimental points are obtained by plotting positive ion current passing through the exit slit of the analyzer as a function of the energy of the ions, while the magnetic field is held constant. As there is a definite relation between the energy and the mass of ions collected, namely, energy $\times$ mass $=$ a constant, the voltage scale can be readily changed to a mass scale, as was done in the figure. The width of the peaks is, of course, caused by the finite width of the slits used and the imperfections in focusing. (After Nier, 1939.)

and lead production; therefore it must disintegrate nearly six times as fast as the other components. Long ago there was a much greater proportion of it in the uranium than there is even now, and it furnished a larger proportion of the resulting lead. Thus, as Nier's table shows, after allowing for ordinary lead, there should be more Pb^{207} in proportion to the Pb^{206} in the older minerals than in the younger (Fig. 55).

Rose made curves showing what the ratio of radium G to actinium D (as the two isotopes Pb^{206} and Pb^{207} are called) ought to be according to two or three different hypotheses as to the rates of disintegration of uranium I and actino-uranium. A. von Grosse had led the way by first making such a curve. Nier has been able recently to get very accurate results and has drawn a curve based on some twenty different minerals and a rock (Fig. 56).

It is a satisfaction to find that the general time scale agrees with previous results. What appears to be the oldest mineral came from near Winnipeg, where on the Huron claim are found both the thorium mineral, monazite, and a uranium mineral, uraninite. Moreover, the mineral deposit on the Huron claim is so old that even one of the less radioactive elements, rubidium, has been disintegrating long enough for Hahn, Mattauch, and Strassmann to be able to trace its change. One of the rubidium isotopes, Rb^{87}, changes to an isotope of strontium, Sr^{87}. Physicists have not been able to agree on the rate of this change, since it is very slow and takes place only by the expulsion of an electron (a beta particle). However, other observations support the conclusion that the change of rubidium and potassium is so slow that even in the oldest minerals the accumulation of the products of disintegration is barely weighable. Now if the world were indefinitely old, a much larger amount of their products would be found. The oldest minerals date back almost 2,000 million years, but the rocks in which they occur must be older yet, so that it is safe to say that the earth is, at least, nearly 2,000 million years old. Also, the fact that no place has been found where the products of the disintegration of uranium, thorium, rubidium, and

potassium indicate a greater age, suggests that it is not so very much older. Some of the astronomers, with their theory of an expanding universe, would shrink the whole universe down to one fat atom in some such time. And just as it is assumed that the earth is older than any mineral or rock of which it is made

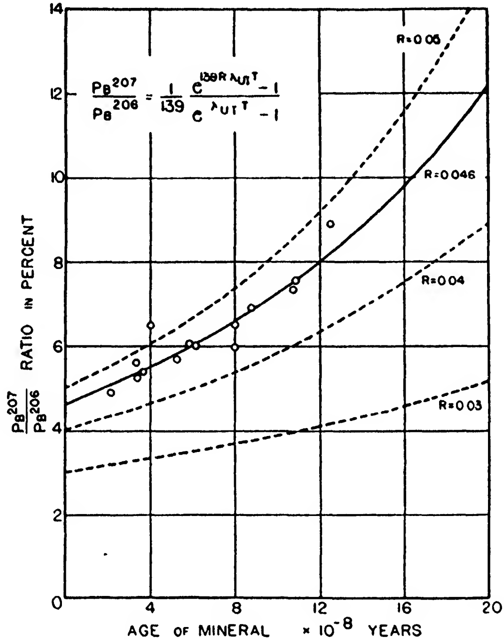


FIG. 56. Age of mineral. Pb^{207}/Pb^{206} (corrected for common lead impurities) as a function of the age of the mineral. The curves were plotted for various assumed values for U^{238}/U^{235} on the hypothetical assumption that the actinium series is just 4 per cent as active as the uranium. (After Nier, 1939.)

up, so it may be assumed that the earth is not older than the universe of which it is a part.

There is one other method of estimating the age of the earth which is as yet undeveloped. Even the "chicken" of hypothesis has hardly cracked its shell. Many of our common elements, such

as hydrogen, oxygen, carbon, nitrogen, are found to be composed of isotopes having slightly different atomic weights. It is not known whether or not they are disintegrating, but Gulbranson and Nier, Urey, Dole, and others have recently found that the proportions do vary slightly in different sources: air, ocean, and organic material. Brewer has also found different proportions of potassium isotopes in organic substances such as bone marrow and plants.

Let us assume that the atmosphere has been deriving a substantial part of its volume from some source like carbon dioxide or volcanic gases, and that some of its elements may be removed by organic or other reactions that tend to remove more of one isotope than another. If these reactions seal up their product as, for example, coal seals up carbon, then there might well be a progressive change in the proportion of the isotopes not thus sealed up but left in the atmosphere or in circulation, and this progressive change might be recorded in products precipitated from the atmosphere.

Gulbranson and Nier, working on the ratio of two carbon isotopes, C^{12} to C^{13}, have obtained some results that suggest this may be true. Birds of a feather flock together, and possibly a tendency for things of a kind to get together may be used in estimating the age of the earth.

Mention has been made of the release of helium "bullets." A natural question is, what becomes of them? Cannot ages be computed by their accumulation as well as by the lead that is formed? This is not so easy to do, however, for helium is a gas and escapes, some of it into the air; and it is a debatable question how much helium there is in the upper part of the atmosphere and whether, under certain conditions, it is not leaking off into space. At any rate, the minute proportion of helium in the atmosphere seems rather too small to agree with the amount of erosion of radioactive rock and leakage of helium therefrom which there would have been in geologic time. But there is helium in the rock and minerals, and, rarely, a well contains a substantial amount.

No minerals that contain radiogenic lead are entirely free from helium, and it was early suggested that their age might be computed from the amount of helium they contain. However, the ages thus computed were soon found to be only fractions—a third or a tenth or less—of the ages that were calculated from the lead in the same minerals. Until it could be certainly established, by the determination of atomic weight and isotopes, that this lead was not ordinary lead, the discrepancies might be explained as due to the presence of ordinary lead; but this explanation is no longer possible. There is no doubt that helium leaks from minerals that contain much of it. Then arose the question whether, if there were only a very little helium in a rock or mineral—to be reckoned in cubic centimeters per ton—it might not be retained, especially in meteorites; for it is found that helium does not leak through iron, though it will leak through ordinary glass. Methods of measuring extremely small quantities of helium were developed. The principle is the same as that which makes an electric light go out when the bulb is broken. A fine wire conducting electricity through a small vacuum tube is very sensitive to the imperfection of the vacuum; a few atoms of helium allowed to leak will make a difference in its conduction of heat and electricity.

Thus Paneth, Urry, Evans, and Goodman, by measuring quantities almost too small to realize—fractions of a cubic centimeter of helium in a ton of rock, and hundred million millionths of a gram of radium in a gram of rock—have studied the accumulation of helium and the radioactivity of meteorites. They have found no meteorites that are more than 2,800 million years old, according to the latest estimates, and it may be assumed that the earth is not much older.

VI

THE EXPANDING UNIVERSE

By H. P. Robertson

Princeton University

> Elpino: How is it possible that the universe be infinite?
> Filoteo: How is it possible that the universe be finite?
> Elpino: Do you mean that one can prove this infinity?
> Filoteo: Do you mean that one can prove this finitude?
> Elpino: What an extravagance of fantasy!
> Filoteo: On the contrary, how narrow your view!

At this point you are doubtless glad to interrupt the opening dialogue of Giordano Bruno's *The Infinite Universe and Its Worlds* with Frascastorio's fervent plea, "*Ad rem, ad rem si juvat!* You've dawdled long enough with empty talk!"—even though, as I hope, you may not already share the hostility of his second auditor Burchio! But Bruno, in the person of Filoteo, goes on to argue a priori for an infinite universe with a plurality of worlds—even as his younger contemporary Kepler was later to argue for the opposite position. In these views Bruno was anticipated by the English mathematician Thomas Digges, whose less metaphysical presentation, incorporated into his account of the Copernican theory, is more likely to appeal to us of a later age; indeed, Digges' quaint diagram illustrating his "perfit description of the Cælestiall Orbes" seems to have been largely influential in grafting these conceptions onto the Copernican system (Fig. 57). But the more judicious Galileo, in keeping with the awakening scientific spirit of the times, refused to be swayed by either metaphysical arguments or inadequate observations, and dismissed the problem for the time being with the considered

opinion that "no one has proved the world to be finite and determinate or infinite and indeterminate."

Bruno's cosmological problem of the structure and population of the universe is as old as thought itself, and one to which the

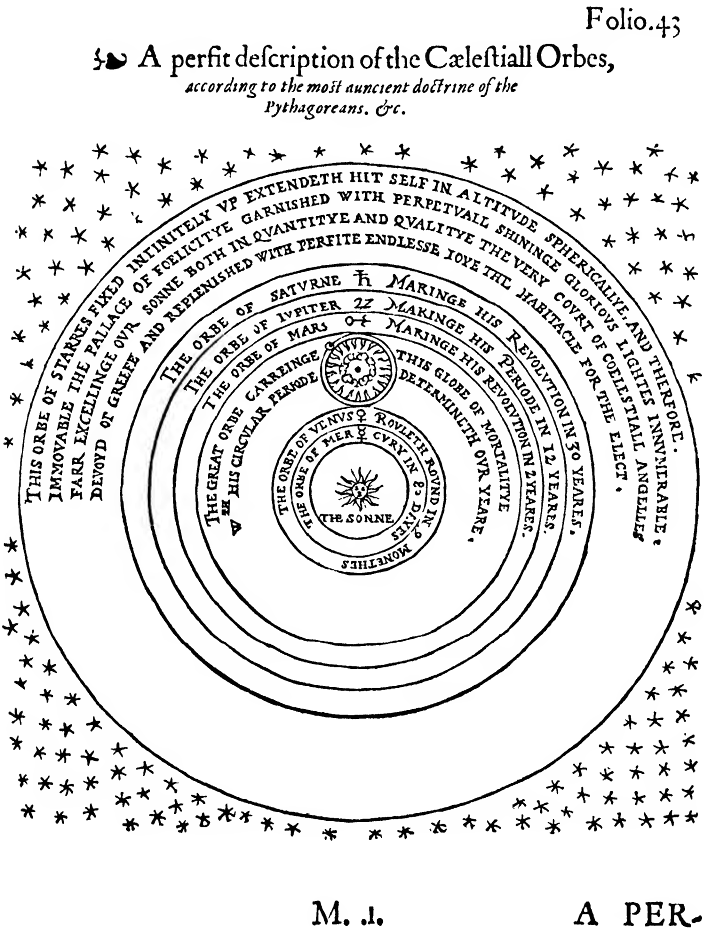


FIG. 57. Thomas Digges' diagram of the Copernican system.

human mind constantly returns. And always the individual approaches the problem under the guidance of some favorite creed —whether it be mysticism or religion, philosophy or science. We, as scientists, looking upon certain aspects of the problem as falling properly within the domain of science, may applaud Galileo's cautious rejection of his contemporaries' bold speculations as not proven by the evidence then at hand. But the two-and-a-half centuries which have elapsed since Galileo's time have witnessed such an advance of the frontiers of knowledge that many problems, then accessible only to other modes of thought, have now been brought within the expanding sphere of the scientific methodology. Indeed, the empirical knowledge of the heavenly bodies amassed within these times, hand in hand with the development of the appropriate theoretical tools, has raised the hope that we may now be within striking distance of the central problems of cosmology.

I hope to show here how far we have proceeded on the way toward such a solution. My course will be charted between the twin guiding lights of theory and observation; may we be blinded by neither the one nor the other, for the way to true understanding must lie between the dangerous rocks of unbridled speculation on the one side, and the shoals of senseless amassing of raw data on the other! We shall be concerned largely with questions concerning the position, distance, and motion of heavenly bodies. Our theories will therefore fall into the general fields of geometry and dynamics, and our observations will be drawn from geodesy and astronomy, largely from the tremendous advances made in this latter field during the past two decades. Throughout it will be necessary to distinguish sharply between geometry and kinematics on the one hand, as mathematical disciplines arising logically from given postulates, and on the other the physical theories of matter and motion resulting from the application, with the indispensable aid of experiment and observation, of one or another of many mathematically consistent systems to the actual world.

The distinction between pure geometry as a mathematical system and applied geometry as a physical theory is well illustrated in the history of geometry from its rise out of the practical rules of Egyptian land surveyors to the development, during the past century, of the so-called non-Euclidean geometries. First, in the hands of the Greeks, geometry became a formal body of mathematical theorems derived from a set of postulates which were, in point of fact, abstracted from the practical surveying rules. The notion of *congruence*, so pivotal in the constructions of Euclidean geometry, is embodied in certain of these postulates as a reflection of the fact that at least for practical purposes the metrical properties of a physical body are independent of its absolute position or orientation in space. Repeated but unsuccessful attempts were made to derive the rather anomalous parallel postulate, which seemed indicated by practical experience, from the simpler postulates dealing with congruences; the failure of these attempts gave rise in the end to the belief that a consistent scheme could be developed without necessarily imposing the parallel postulate.

During the first half of the nineteenth century this belief matured into achievement at the hands of the great mathematician Gauss and his gifted younger contemporaries Bolyai, Lobachewski, and Riemann. The least strange of these geometries, that due to Riemann, is one in which the analogue of plane Euclidean geometry may be taken as the geometry of the surface of a sphere of some definite radius R. The role played by the infinite straight lines of the Euclidean plane is here played by the great circles of the sphere; hence each full line on the sphere is closed and of finite length $2\pi R$. Since each great circle intersects every other one, there are no parallel lines; further, the sum of the interior angles of a triangle on the sphere is greater than two right angles, and by an amount which is proportional to the area of the triangle. However, the central notion of congruence is preserved in this "spherical" geometry, for each configuration of arcs of great circles may be reproduced, without change of its

intrinsic metrical properties, in any position or orientation on the sphere.

Clearly the *radius* of the sphere is not itself an element of this geometry, and should therefore be replaced by some notion intrinsically associated with the surface; the most convenient concept is that of the *curvature* K of the surface, which is defined numerically as equal to the inverse square $1/R^2$ of the radius R. Because of the really great importance of this notion for our eventual conclusions, I hope you will bear with me while I show that this curvature can, in fact, be determined by measurements performed on a limited portion of the sphere, for upon such considerations depends, in the last analysis, the possibility of determining the "radius of curvature," or better the "curvature," of the physical universe. I must therefore ask you to consider a small circle on the sphere, whose radius *on the sphere* is an arc of length r of a great circle of the sphere. Now it should be apparent to you that the perimeter p and the area a on the sphere of this small circle are, because of the curvature of the sphere, actually *less* than the perimeter and area of a circle of the same radius r in a Euclidean plane; the deficiency is in fact a measure of the curvature in the technical sense defined above.[1] In this sense, the determination of curvature implies no mystical ability to crawl outside of the space—in this case the spherical surface—whose curvature is to be measured. And since the units in which distances are measured are more familiar than those of curvature, it will be found easier, perhaps for psychological reasons, to translate curvature over into a hypothetical "radius of curvature" $R = 1/\sqrt{K}$, even though we confine ourselves to measurements performed within the actual space.

Let us return from this rather pedantic analysis of the two-dimensional spherical geometry to the full three-dimensional

1. I shall not trouble you with the exact formulae for these expressions; the following approximate ones are sufficient for our purpose, so long as r is not too large a fraction of the distance $2\pi R$:

$$p = 2\pi r(1 - Kr^2/6), \quad a = \pi r^2(1 - Kr^2/12).$$

"spherical space" of curvature K. Here each "straight" line is of finite length, without it being necessary for us to think of it as curved in some other dimension, and each two intersecting lines determine a two-dimensional surface whose intrinsic geometry is the same as that of the surface of a sphere of curvature K. In short, we have here an internally consistent congruence geometry in which space is unbounded—there are no boundaries past which we cannot go—but of finite volume.

The third congruence geometry, discovered independently by the three other mathematicians mentioned above, is one in which each " straight" line is infinite in length, and in which there exist an infinite number of parallels to a given line. In this "hyperbolic" geometry the sum of the three interior angles of a triangle is *less* than two right angles, and the perimeter and area of a small circle *exceed* those of a corresponding Euclidean circle—and by an amount which is given by the formulae set forth above, provided we take the curvature K appearing therein as a *negative* number. A two-dimensional section of this hyperbolic space has a geometry which is intrinsically the same as a certain spindle-shaped surface, called the "pseudo-sphere," which may be erected in Euclidean space.

There are, then, three mathematically consistent congruence geometries: (a) the spherical (or, differing slightly from it, the elliptic) geometry, with a positive curvature K; (b) the Euclidean (or parabolic) geometry, for which $K = 0$; and (c) the hyperbolic geometry, in which the curvature K is negative; further, the later work of the Norwegian mathematician, Sophus Lie, shows that these three are the only possible congruence geometries. Which, then, if any of the three, is applicable to physical space? As Helmholtz noted in *Concerning the Facts Which Lie at the Basis of Geometry*, the assumption of the existence of rigid bodies, which may be moved about in physical space without change of their intrinsic metrical properties, requires that physical geometry be a congruence geometry. Because of the approximate validity of Euclidean geometry in terrestrial affairs

—Gauss himself, in connection with a geodetic survey of his native province of Hanover, had remarked on its validity there to within the errors of observation—it is clear that the empirical distinction between the three possibilities can most hopefully be sought in the vast distances met in astronomy.

A serious attempt at this determination was made by K. Schwarzschild at the turn of the century. In keeping with standard astronomical practice, he postulated that "a triangle between three points shall be defined as the paths of beams of light from each point to each other, the lengths of the sides by the times required for light to transverse them, and the angles shall be measured with the usual astronomical instruments." It had long been noted that a negative parallax established for even a single star would imply spherical space, but Schwarzschild set himself the more modest task of determining the permissible curvatures for the two non-Euclidean geometries. In his own words: "One there finds oneself, if one but will, in a geometrical fairyland, but the beauty of this fairy tale is that one does not know but that it may come true. We accordingly discuss the question of how far we must push back the boundaries of this fairyland—of how small we must choose the curvature of space, or how great its radius of curvature." Schwarzschild's results, deduced from data on the nearer stars of our own system, can hardly be expected to yield very startling conclusions concerning the immensity of space, but they are significant as an earnest of his clear recognition of the empirical foundations of physical geometry; in brief, he found that on the hypothesis that space is elliptic its equivalent radius $R = 1/\sqrt{K}$ must be greater than 1,600 light-years, and that the corresponding quantity $R = 1/\sqrt{-K^2}$ (a distance in which curvature effects build up to a certain well-defined amount) for hyperbolic space must exceed 63 light-years! [2]

Before passing to a discussion of the role of curvature in physical theory, let us dwell for a moment on the apparent discrep-

2. One light-year, the distance which light will travel in one year, is approximately 6 million million miles.

ancy between the position maintained above and that of the late great French mathematician and natural philosopher, H. Poincaré—an examination which is all the more timely in that Poincaré's position has of late been distorted and misapplied by the English astrophysicist, E. A. Milne, in his treatment of the cosmological problem. Poincaré did hold, both in his *Science and Hypothesis* and in his *Science and Method*, that no experiment could possibly either prove or disprove the validity of either the Euclidean or non-Euclidean postulates. Thus, in the opening of the chapter on "Experience and Geometry" in *Science and Hypothesis*, Poincaré states: "If therefore negative parallaxes were found, or if it were demonstrated that all parallaxes are superior to a certain limit, two courses would be open to us; we might either renounce Euclidean geometry, or else modify the laws of optics and suppose that light does not travel rigorously in a straight line." He continues: "It is needless to add that all the world would regard the latter solution as the more advantageous." It is perhaps needless to add that, in this, Poincaré was almost as bad a prophet as his cousin was a statesman, for at present almost all the civilized world does accept the alternative solution, and regards the "bending" of light around the sun as evidence of the straightness of its path in a non-Euclidean geometry! But to follow Poincaré in this would be but to beg the question. I am not worried, with him, as to whether "any concrete magnitude which I have measured with my material instrument *really* represents the *abstract* distance" (italics mine); I am only concerned, as is Schwarzschild, with the question: *Given* a concrete procedure for the measurement of distance and angles, is the *physical* geometry which this procedure reveals Euclidean or non-Euclidean? Perhaps Poincaré was at worst a bad prophet; I cannot for an instant believe that so clear a thinker would have sanctioned a cosmology, based on the assumptions of Euclidean space *and* the rectilinear propagation of light, as anything more than a provisional hypothesis—and certainly not as an a priori form to which the physical world *must* conform.

Such is the problem of the empirical determination of the curvature K of physical space; but there remains the deeper question of *why* any one numerical value of the curvature rather than any other—the type of question which is meaningful in science only if it can be attacked in terms of the correlations predicted by a physical theory. The possibility that the metrical relations of space are determined by the physical properties of its contents was envisaged by Riemann in his celebrated inaugural address *Concerning the Hypotheses Which Lie at the Basis of Geometry*, but its concrete expression was only realized fifty years later in Einstein's general theory of relativity. According to it, the natural geometry of space is determined by its material content; crudely put, the presence of a bit of matter near the region under consideration influences the curvature of that region, in a manner defined by Einstein's equations of gravitation.

Now in its application to the actual world, Einstein's theory does not in general lead to a congruence geometry; the geometry of a region is determined by its position relative to matter, and we can, therefore, no longer assume that all physical properties of space are independent of position. Thus the curvature at a point in the solar system will vary in accordance with its distances from the sun and planets, but in a way which is determinate in terms of the distance to and the masses and velocities of these bodies. If, however, we are interested primarily in the geometry of huge portions of the universe, and if we find that matter is strewn about in a more or less random fashion, it would seem legitimate to ask what the average curvature of such a region would be—much as we can determine a sensible average curvature over a portion of the earth's surface, in spite of its obvious topographical irregularities. Since these assumptions are at least approximately satisfied in the portion of the universe accessible to the largest existing telescopes, and assuming that this represents a fair sample of the universe, we may reasonably expect that Einstein's theory will lead us back to one of the congruence geometries.

To find the first attempt to determine the natural geometry of the universe along these lines, we must go back twenty years to the solution proposed by Einstein himself. At that time there was no cogent reason for believing that the great galaxies of stars outside our own Milky Way system were in any systematic motion relative to each other, and small random motions would not be expected to affect materially the program above proposed. In short, it then seemed reasonable to assume that the properties of space were not to be considered as varying systematically with the time of observation, and that Einstein's field equations could be used for the determination of the curvature due to an essentially uniform and static distribution of matter. Neglecting the undoubtedly small pressure effects (due to radiation pressure or the random motions of matter), Einstein found that the curvature of space is directly proportional to the mean density ρ of matter, with a positive constant of proportionality which depends only on universal physical constants.[3] This "Einstein universe," as it has come to be called, is accordingly one in which space has a finite total volume, and in which the equivalent constant radius R turns out to be proportional to the total mass M of all matter in the universe.

Having thus shown how the empirical evidence, augmented by extrapolation and on the basis of Einstein's theory of gravitation, leads to a definite natural geometry for the universe, we next turn to astronomy for the necessary numerical data. According to the estimates of E. P. Hubble, the mean density of the matter contained in the nebulae, when smoothed out over sufficiently large portions of space, is at least 10^{-30} grams per cubic centimeter, and may be as much as 10^{-28}. On substituting this first result into Einstein's formulae above, we find that the curvature is defined by an equivalent radius R of 35,000 million light-years—so large that it would take light some 200,000

3. The precise results are

$$K = 4\pi G\rho/c^2, \quad R = 2\,GM/\pi c^2,$$

where G is the Newtonian constant of gravitation, and c is the velocity of light.

million years to circumnavigate it. And large enough to contain many millions of millions of nebulae, each consisting of 100 million suns—astronomical figures with a vengeance! The larger density limit only reduces these figures to one tenth as much. Here, then, is a possible answer to Bruno's question—a finite universe, in a sense in which Bruno did not dream, and containing a finite number of worlds.

But doubt was soon to be cast on the tenability of this solution. For shortly before this V. M. Slipher, of the Lowell Observatory at Flagstaff, had observed the Doppler shift in the spectrum of the light received from a nebula, and by 1923 had measured this shift for a total of forty-one nebulae. In all but five of these nebulae the Frauenhofer lines were shifted toward the red end of the spectrum—a result which, taken at face value, would seem to indicate that the great majority of these nebulae were moving away from us. The exceptions were for the most part the brighter, and therefore presumably the nearer, nebulae—whereas the fainter nebulae were all moving away, with velocities which seemed to increase with their faintness. The continuation of this work up to the present, principally by Hubble and Humason at Mount Wilson Observatory, has increased the number of nebulae observed for this effect to around two hundred, and has amply confirmed the first vague intimations that the fainter the nebula, the greater the red-shift. Hubble's careful analysis of this rich material has established an empirical correlation between red-shift, or its equivalent velocity, and distance; the velocities increase linearly with distance, the increase of velocity per million light-years in distance being very nearly 100 miles per second —empirically up to a velocity of 25,000 miles per second, observed in light from a nebula whose distance is estimated at 240 million light-years! The deviations of the actual radial velocities from this relationship, which average little over 100 miles a second and seem not to be correlated with distance, are attributed to random motions; when we remember that the only significant absolute velocity in our theories is the velocity of light, it would

seem that the motion of a nebula is essentially prescribed by the region of space in which it finds itself—assuming that the transverse velocities, on which we have no evidence, would not belie this picture. It should be remarked that we have *inferred* that the red-shift is a true Doppler effect due to radial motion, but that we have no other direct evidence for this interpretation. However, we *know* of no other way in which shifts of such magnitudes could be effected, and while it must be allowed that we may here be dealing with some cause which has so far escaped detection in other connections, we are acting in accord with good scientific methodology in pursuing the consequences of the inference that the shift is due to motion—until we are faced with some insuperable obstacle to this interpretation, or until other evidence reveals an alternative mechanism for its production.

How does this new-found situation affect our search for a physical geometry? In the first place, we have here a new element, or at least one whose effects could until now be largely ignored, namely, the recognition of large systematic motions in the matter responsible for the curvature. The static Einstein solution is no longer tenable, for it was derived on the assumption that no such motions existed. But before proceeding to the solution of the field equations under these altered circumstances, let us attempt qualitatively to foresee the direction the answer will take. The relativity theory of gravitation attributes the curvature properties of space to the proximity of masses; if, therefore, the masses are receding we should expect the curvature to decrease, and hence the length R to increase. That is, the curvature of our congruence geometry may change with time—a suggestion put forward half a century ago by the English philosopher-mathematician, W. K. Clifford, and by the French philosopher, A. Calinon! Had they the data now in our possession, they might well have concluded that the observed motions of the nebulae were a direct concomitant of an increasing radius of curvature of space—much as the spots on a speckled toy balloon recede from one another when it is inflated. But we must go further,

for the field equations should tell us, as well as the magnitude of the curvature, its present and past history, provided only we have enough data concerning the present epoch.

The existence of these motions compels us to take explicit account of a fact which we have so far tacitly ignored: the fact that the theory of relativity is based on a four-dimensional geometry, in which time plays the role of the fourth dimension. The three dimensional geometry, designed to deal with distances and orientations, must be augmented to include the kinematical elements of duration and velocity. The problem of a congruence geometry in this extended sense can still be set and solved; whereas formerly it consisted in determining the metrical structure of a space in which each point and each direction is equivalent to each other point and direction, it now consists in determining the structure of a space time in which in addition each epoch and each velocity is equivalent to each other epoch and velocity (provided the velocities are both less, or both greater, than the critical light velocity). The simplest congruence geometry in this sense, the analogue of that of Euclid, is the space-time geometry of Minkowski implied by the special theory of relativity; in it one may speak of parallel motions, as well as of parallel directions. Of the remaining types, the simplest is that first discussed by the Dutch astronomer, De Sitter, as an alternative to the static Einstein cosmology. But although this "De Sitter universe" was partially successful in describing the motions of the nebulae, it, in common with all congruence space times, is basically in conflict with the empirical fact that all motions are *not* equivalent, for the observations indicate that there is a preferred state of motion associated with each region of space, namely, the motion of the nebulae actually occupying the region in question. The nebulae are *not* swarming helter-skelter past each other, with velocities ranging up to that of light, as a congruence geometry would imply.

What is required is, in fact, a solution very similar to that sketched on page 158; a space time in which at each epoch t the

three-dimensional space is a congruence space having a curvature $K(t)$ which changes from time to time. This, in a sense, reintroduces a universal time t, in keeping with the fact that there is a preferred state of motion associated with each point of space. The laws governing the change of K in time and relating it to the empirical data are then obtained with the aid of Einstein's equations of gravitation, as shown by the work of A. Friedmann, G. Lemaître, R. C. Tolman, and myself. This is hardly the occasion for a detailed account of these highly technical developments, and I shall therefore confine myself to a brief statement of the conclusions which may with reasonable safety be drawn from this work.

First, what is it that is expanding (or contracting, as the case may be) in such model—are we to think of the points of space as flying apart in some mysterious way, and dragging along with them the particles of matter which happen to occupy them at the moment? For if so, how could such a motion be observed? Our eyes, our measuring rods, our telescopes and all the world would be expanding merrily at the same specific rate, and we would be completely unable to detect the recession of the nebulae —even as Poincaré found nothing amiss on awakening after a night in which, he supposed, everything including his own body had increased to a thousandfold its normal size! The answer the field equations offer to this question is that any subsystem which is bound together by the mutual gravitational or electrical forces of its parts does *not* partake of the expansion; a measuring rod, a solar system, a double star, a galaxy, or even a cluster of galaxies so massive or so close together that their motions are dominated by the gravitational attraction of their fellows, are all unaffected in their internal metrical relations. Our homely illustration of the speckled balloon is faulty if the specks expand during the inflation; we should instead think of them as flexible, but inelastic, patches pasted on the surface.

In order to determine the whole course of the curvature $K(t)$, it is necessary to have three observational data concerning

the present epoch of the universe; but the evidence touched upon so far in this discussion yields only two, the present mean density of matter and the constant of proportionality in the velocity-distance relation. For the remaining requisite datum either of the following two would suffice: an accurate determination of the systematic deviations from the linear velocity-distance relation at greater distances, or a more or less direct determination of the present value of the curvature K. The former presumably must await the construction of more powerful instruments for spectrographic research—perhaps the 200-inch reflector now under construction at the California Institute of Technology will bring this possibility nearer realization. A determined attack on the alternative possibility, the present value of the curvature, has already been made by Hubble in his extremely difficult and highly ingenious work on the numbers of nebulae observed out to faint apparent magnitudes, the story of which has been clearly and delightfully told in his Silliman lectures, *The Realm of the Nebulae*.[4]

I hope you will again bear with me while I attempt to show how this curvature might, in principle at least, be determined from the observational material. It has indicated previously (page 151) that the perimeter and area of a small circle, of radius r measured along a "straight" line of the space, would differ from the corresponding quantities of a Euclidean circle of radius r by amounts which would depend upon the curvature. Similarly, the area A and volume V of a small sphere of radius r, constructed in a space of constant curvature K, will differ from the corresponding quantities in Euclidean space by amounts depending on K, the former being again systematically less than the latter if K is positive, and greater if K is negative.[5] Hence if we have reason to believe that the nebulae are scattered more or less uniformly throughout the observable portion of the universe, the

4. Yale University Press, 1936.

5. The approximate formulae are here

$$A = 4\pi r^2(1 - Kr^2/3), \quad V = \tfrac{4}{3}\pi r^3(1 - Kr^2/5).$$

number of nebulae within a distance r of the observer will increase in direct proportion with the volume of a sphere of radius r, and the deviations of these results from the Euclidean behavior should yield a value for the curvature.[6]

The data on which Hubble's determination is based consist of five photographic surveys of regions of the sky, taken with the 60-inch and 100-inch reflectors at Mount Wilson and by Mayall with the 36-inch reflector at Lick Observatory, under differing exposure conditions, and encompassing in all more than forty thousand nebulae. Each such survey yields the number per square degree of nebulae brighter than some determinate limiting apparent magnitude m, ranging for the various surveys from $m = 18.5$ out to $m = 21.0$—the latter limit corresponding to nebulae one-millionth as bright as the faintest star visible to the naked eye, nebulae whose distance from us is estimated at some 400 million light-years! Before these results can be compared with theory they must be corrected for a great many factors, among others for obscuration by cosmic dust within our own galaxy, and above all for the weakening effect of the red-shift itself. Indeed, the red-shift causes a nebula to appear fainter than it would otherwise appear for two reasons, (*a*) the individual photons carry less energy, in accordance with Einstein's law asserting the proportionality of energy with frequency; and (*b*) since

6. The actual situation is somewhat complicated by the fact that in practice the distance of a nebula is inferred from its observed luminosity. It follows from the optical principles involved that the "optical distance" d so measured is not in fact the distance r which would be measured by more direct means, but is rather the same as the radius of a Euclidean sphere whose area is equal to the area A of the non-Euclidean sphere considered above. On setting the expression for A equal to $4\pi d^2$, we find to the approximation here carried that

$$r = d(1 + Kd^2/6),\ A = 4\pi d^2,\ V = \tfrac{4}{3}\pi d^3(1 + 3Kd^2/10);$$

hence the volume of a non-Euclidean sphere of optical radius d is greater than that of a corresponding Euclidean sphere if K is positive, and less if K is negative. Since Hubble's observations indicate that, after allowing for all known relevant factors except curvature, the number of nebulae observed increases faster with distance than would be expected in Euclidean space, he is led to the conclusion that the curvature of physical space is positive, i. e., that it is a closed space of the spherical or elliptical variety.

on the theory we are here considering the nebulae are actually receding from us, we therefore receive from them fewer photons per second than we would if they were stationary. This second correction turns out to be a matter of great importance in the interpretation, for if we should assume some mechanism, other

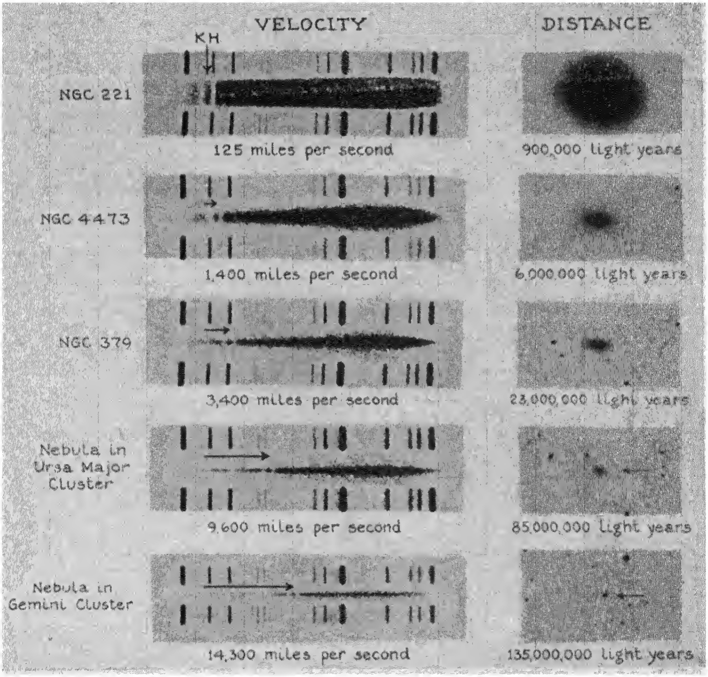


FIG. 58. Illustrating the evidence upon which the empirical velocity-distance relation is based. The arrows above the nebular spectra point to the H and K lines of calcium and show the amounts these lines are displaced toward the red end of the spectra. The comparison spectra are of helium. The direct photographs (on the same scale and with approximately the same exposure times) illustrate the decrease in size and brightness with increasing velocity or red-shift. NGC 4473 is a member of the Virgo cluster and NGC 379 is a member of a group of nebulae in Pisces. (From spectra and photographs obtained by Hubble and Humason at the Mount Wilson Observatory.)

than actual velocity, at work in causing the shift, we might need only to apply the first of these two red-shift corrections (Fig. 58).

To make a long and complicated story short and simple, on correcting the apparent magnitudes, and hence the distances, for only the first of these two factors, the data are found to be consistent with the assumption that the nebulae are distributed at random over a portion of space in which the curvature effects are negligible, for the numbers of nebulae then increase with the cube of the distance, as they should if space were Euclidean. But on applying also the second red-shift correction to the counts—and we *must* if we assume that we are here dealing with a true velocity of recession—Hubble finds that it is then necessary to introduce a compensating correction which, if attributed to the effects of curvature, corresponds to a certain definite large positive curvature of space.

Here, then, is apparently the additional datum required in order that we may complete the solution of the equations which govern the behavior of K as a function of the time, and thus to answer questions concerning the remote past or future of the universe. But at the outset one startling fact emerges; Hubble's present value of K is so large that the universe must be much smaller than might otherwise have been expected—the equivalent radius R is at present no more than 500 million light-years, corresponding to a spherical universe whose total "circumference" is but 3,000 million years (or half as much if we assume elliptic space). And Hubble has photographed nebulae whose inferred distance is 500 million light-years, a very respectable fraction of the total distance around! So far, however, we are not actually in conflict with other known facts, but an unpleasant surprise awaits us when the theoretical expression for the mean density of matter is compared with the observations. For the curvature is so great that unless we are willing to assume an inordinately large deviation from the essentially linear velocity-distance relation, just beginning at the limit of present observations, the computed density turns out to be enormous—sixty times

as great as Hubble's *maximum* estimate of the smoothed-out density of the matter contained in the nebulae. Even this is—just barely—conceivable, for there *may* be enough dark matter strewn about in space to warrant such a figure; but here we are already treading on questionable ground, for although such an assumption would not be as contrary to the canons of scientific procedure as the introduction of a totally new process to explain the red-shifts, it is nevertheless distinctly disquieting. But worse awaits us, because if we are willing to accept such a density and proceed to the integration of the equations satisfied by the curvature, it is found that they imply that less than 1,000 million years ago the curvature was so great, and hence the universe so small, that all the nebulae must then have been practically mixed up together. It is not this latter aspect which is so disturbing, since on other grounds it is even plausible that the present stage evolved from some such state as here depicted, but rather the untenably short-time scale which is thus allowed for the evolution of the nebular system. Indeed, the time required for the earth to evolve from a stage in which rocks were already formed is greater than this, for among such rocks are found accumulations of radioactive by-products which could only be amassed in something like 2,000 million years—unless, again, we are willing to espouse another wild hypothesis which is in the air, and assume that the rate of radioactive processes was accelerated when the world was young!

We seem to have come to an impasse with this line of attack; I find myself unable to accept the model to which so large a curvature almost inevitably leads, and unwilling to postulate *ad hoc* some new principle to lift myself over the difficulty. Looking back over the assumptions involved in the determination of the present value of the curvature, we find that at many points the conclusion is sensitive to even relatively slight uncertainties; above all, the method is highly sensitive to lack of uniformity in the distribution of the nebulae, whether this be due to fluctuations in the statistical material or to large-scale structural features of the nebular system. That density gradients, of such a magnitude

as to raise considerable doubt concerning the validity of the method, do in fact exist in regions closer to us than those emphasized in Hubble's surveys, is clearly shown by the recent survey by Shapley of some 75,000 nebulae brighter than the eighteenth magnitude in southern galactic latitudes. On arbitrarily dividing all his material into the Eastern and Western hemispheres, Shapley finds that the number of nebulae per square degree in the former exceeds that in the latter by between 40 and 50 per cent; similarly, he finds that the number of nebulae per square degree in the central portion of the plates taken in the southeast quadrant is more than twice the corresponding number for the northwest quadrant, and it is to be noted that he is even here dealing with a total of almost 20,000 nebulae. The existence of similar density gradients along the line of sight at the distances involved in Hubble's surveys might easily mask or distort the curvature effects which we are seeking. Until more positive evidence on this point is available, it would therefore seem expedient to seek our additional datum elsewhere—perhaps most hopefully in an accurate determination of the deviations from the linear velocity-distance law at greater distances, for this effect should be much less sensitive to departures from the distribution implied by our homogeneous models.

Throughout the discussion the emphasis has been placed on the oversimplified models which lead to a more or less unique stratification of space time into space and time, and in which the geometry of space is one of the congruence geometries. Once we have found such a homogeneous metrical background for the universe, the field equations will permit the computation of the deviations from this idealized solution which are due to the observed irregularities in the distribution and motions of matter. But I hope that my treatment has made it at least plausible that the problem of the geometry of physical space is a physical one, and one which, in the present state of natural knowledge, can most hopefully be attacked through the field equations of the general theory of relativity. From this standpoint even the dis-

covery of an alternative mechanism for the production of the red-shift would not alter the fundamental problem, since, even though the shift were not attributed to a changing curvature, we would still be faced with the problem of determining the underlying physical geometry. And among the great wealth of possible geometries that of Euclid plays no overwhelming role; apart from mathematical or psychological attributes, which should scarcely be of great import for physics, it is merely one of the already highly specialized congruence geometries, and why any one value of the curvature should be singled out above any other is itself a physical question which requires a physical answer.

Here, at any rate, is Bruno's problem in modern form, together with a brief account of the kind of answer which physical science may hope to give. Forgive me if I may have seemed too partisan a Filoteo; I had no need of that, for here we are happily without a Burchio to exclaim, "I won't believe it, no matter how true it is!"—and then to leave before the discourse is at its end!

VII

COSMIC RAYS AND NEW ELEMENTARY PARTICLES OF MATTER

By Carl D. Anderson

California Institute of Technology

Although research into cosmic rays has actually been in progress for some forty years, these researches at present are being pursued with an intensity and vigor greater than at any previous period in their history. In scientific research, in general, one finds that the solution of one problem will usually reveal the presence of one or more new problems which are often more fundamental in character. This tendency of a scientific field of investigation to become increasingly complex as it develops, is well exemplified in the investigations of cosmic rays. At present the questions of the greatest interest and importance which have been raised in the course of cosmic-ray studies are those which are as yet unanswered.

The work of the past forty years has served to show very clearly the intimate association between the phenomena of cosmic rays and those facts which are basic to our understanding of the structure and properties of matter. Cosmic-ray investigations have already resulted in the discovery of the positron, one of the four well-known basic constituents of matter. More recently strong evidence has been found for the existence of still another type of fundamental particle of matter. These particles, called mesotrons, seem to have a mass intermediate between that of the proton and the electrons, and so far have not been encountered in any experiments except those on cosmic rays. New and quite

novel types of atomic disintegrations have been observed. These differ in important respects from the disintegrations produced in the laboratory by means of a cyclotron, or one of the other high-voltage machines now employed in the study of atomic disintegrations. The differences can readily be explained by the fact that the highest energy particles, which so far have been produced directly by these laboratory devices, have energies under 20,000,000 electron volts, whereas particles occur in cosmic rays whose energies by direct measurement have been found to exceed one thousand times the above value. Cosmic rays then at present provide the only source of particles which are imbued with energies of such prodigious magnitudes. In the course of cosmic-ray work in the future many important new facts will certainly be uncovered when the capabilities of these very high energy particles in producing atomic disintegrations have been adequately explored. In this connection, however, cosmic-ray researches suffer from a severe limitation. In the various laboratory devices for producing disintegrations, it is a relatively simple matter to accelerate particles to high energies in such numbers that several billion particles will strike the target under investigation in every second. A comparison of this with the number of cosmic-ray particles available, which at sea level amounts to only one per square centimeter per minute, will show why the study of disintegrations produced by cosmic rays is in comparison such a slow and tedious process.

In addition to serving as an important tool for the physicist in the study of atomic structure, cosmic rays are a potential tool for the use of the astrophysicist in his studies of the extraterrestrial phenomena of the universe. Investigations have shown that cosmic rays have their origin somewhere in the remote regions of space, but just where the regions are, or what the events are, which give rise to cosmic rays is pretty much a mystery at the present time. The usefulness of cosmic rays as a tool of astrophysical research will therefore have to wait until the knowledge concerning their mode and place of origin is placed on a firm

foundation. This can come only after the completion of prolonged studies of the manifold properties of the rays themselves and of those phenomena produced when the rays are absorbed in their passage through matter.

METHODS OF DETECTION

The very earliest cosmic-ray researches at the opening of the present century were concerned largely with showing merely the presence of a new radiation. The apparatus used at that time was able to do little more than indicate the presence of the radiation, and to measure only one of its important characteristics, namely, its penetrating power, which is a measure of the ability of the rays to pass through various thicknesses of different substances.

The Electroscope

Cosmic rays, of course, are not directly discernible by any of our physical senses, they are too weak to affect a photographic plate, and the only way they can be detected at all is indirectly through their electrical properties. In principle the detecting device, the electroscope, is very simple. An electroscope consists merely of a small body, electrically insulated from its surroundings, upon which a small charge of electricity can be placed, and of a device for measuring the amount of this electric charge at any instant of time.

One type of electroscope is shown schematically in Figure 59. The central metal rod which passes through an insulating cork in the top of the container has two pieces of gold leaf fastened to its lower end. When an electric charge is placed on this rod, the gold leaves will also become charged and the mutual repulsion of these charges will cause the gold leaves to separate and stand apart as indicated in the figure. The degree of separation of the gold leaves is, then, a measure of the amount of electricity which has been placed upon the metal rod. If now, for one reason or another, this amount of electricity should gradually decrease,

the gold leaves would gradually approach one another, and the rate of decrease of the electric charge could be measured by the rate of fall of the gold leaves.

Cosmic rays are able to discharge such an electroscope, and the rate of discharge is a measure of the intensity of the radiation. The way in which this discharge comes about is also indicated in Figure 59. The space within the container is filled with air or some other gas. If now a high-speed cosmic-ray particle, as for example an electron, passes through this air it will separate electric charges from certain of the molecules of the gas along its path. There will then be present in the gas a number of charged molecules, or *ions*. The ions which bear a negative charge will be attracted by the positively charged central rod, and upon being drawn up to it and striking it, will neutralize part of the charge it carries and cause the gold leaves to converge slowly. Thus an electroscope can serve to measure the total intensity of the cosmic rays at any position on the earth's surface.

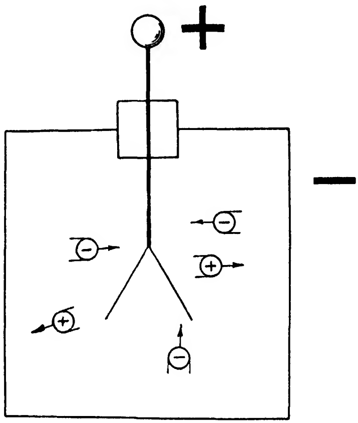


FIG. 59. Electroscope and ions. Schematic diagram indicating positive and negative ions of the gas.

The electroscopes used in the early cosmic-ray experiments were relatively crude in their construction and not highly sensitive. Constant development, however, has improved the instrument until today several types of electroscope have been produced which are highly sensitive and completely automatic in their operation. A continuous record of cosmic-ray intensity is gathered and imprinted on photographic films. The electroscope recharges itself at regular intervals, and requires no attention for days at a

time except the occasional winding of the clock mechanism which acts as a driving motor (Fig. 60).

An important part of the general program of cosmic-ray research has been the systematic study of cosmic-ray intensities

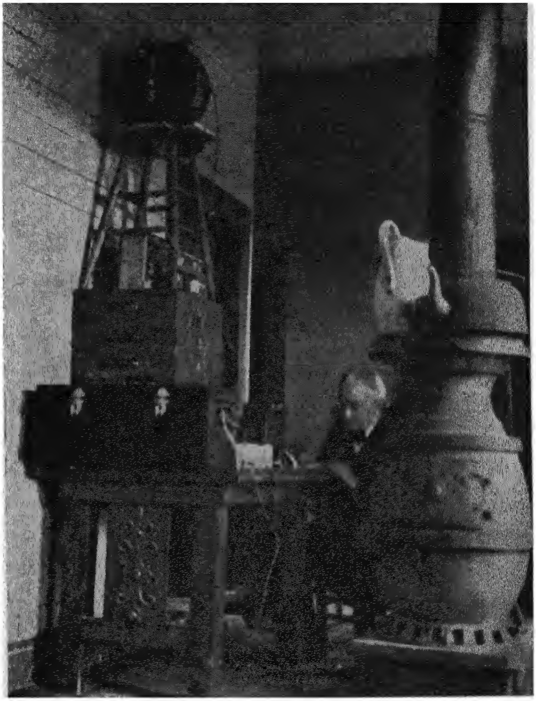


FIG. 60. A cosmic-ray electroscope incased in a spherical lead shield in position in a building on the summit of Pike's Peak. The electroscope is placed near the ceiling of the room in order that radiations from the surrounding soil and rocks may be reduced to a minimum. The observer is Dr. R. A. Millikan.

by means of these instruments when operated at various selected positions on the surface of the earth. Recordings have been made with electroscopes carried into deep mines, and to the tops of

high mountains. Expeditions have carried them to the earth's equatorial regions, into the far north, and the Byrd Expedition

FIG. 61. A string of five balloons carrying a cosmic-ray electroscope, preparatory to a flight in Texas. In order to reach far into the stratosphere as many as ten balloons on one string are often employed.

has made measurements in the Antarctic. By means of manned balloons and free balloons they have been carried above the earth's surface to heights exceeding 90,000 feet above sea level (Fig. 61). They have been lowered into the sea, into the depths of high-altitude mountain lakes. They have made many round-the-world journeys by boat, where they have performed their duties of gathering a continuous record of cosmic-ray intensities, without attention except the daily winding of the clockwork mechanism. Experiments are now under way to develop instruments to be sent up in free balloons in remote regions of the earth's surface where a subsequent recovery of the balloon and instruments is impossible. For this purpose a radio transmitter carried in the balloon is necessary in order that the record of the cosmic-ray intensities can be automatically signaled back to the observers who have remained at the ground station.

We shall see later what some of the results of the electroscope experiments have been but, first, we shall mention two other kinds of experiments which have brought important information about the properties of cosmic rays. These are the cloud chamber and the Geiger counter.

The Cloud Chamber

The development of the cloud chamber had its genesis in one man's consuming love for clouds. Some forty years ago there lived in England a young man who, as a boy, was in the habit of climbing to the tops of hills to watch the clouds float by. Later he became a physicist, but never lost his interest in clouds. He carried out many important studies on the properties of thunder and lightning, for one could not have lightning without clouds. He also made his own clouds in the laboratory, by placing water in a closed vessel and then suddenly expanding the space enclosed by his vessel, thus cooling the air within it and forming a cloud. Then in attempts to measure the electric charge carried by the electron, he would measure the rate at which his clouds fell in the presence of an electric field. But one day when so en-

gaged he made a discovery of the first magnitude. He noticed little wisps or lines of condensed vapor. He was soon able to show that these little lines of vapor represented exactly the paths followed by individual atomic or subatomic particles as, for example, electrons. This experiment, which has been called "the most beautiful experiment in physics," has given to the world one of the most powerful techniques for studying the properties of atomic or subatomic particles. The name of the experimenter was C. T. R. Wilson, and the apparatus has been called after him, the Wilson Cloud Chamber.

The action of a Wilson cloud chamber, as applied to cosmic rays, is as follows. A high-speed cosmic-ray particle as it passes along through any material substance, be it solid, liquid or gas, collides with a number of molecules along its path and in so doing removes electric charges from a number of these molecules, thus converting them into electrically charged molecules, or ions. It is just these ions which are responsible for the action of the electroscope as we have seen previously. But let us now consider this matter from a slightly different point of view.

The high-speed cosmic-ray particle is, of course, itself invisible and so is the trail of ions which it leaves in its wake. If the material through which the particle passes happens to be normal air, about two or three hundred pairs of ions will be formed along each inch of its path. Figure 62 will indicate such a series of positively and negatively charged invisible ions left in the air after the passage of a cosmic-ray particle. The cloud-chamber apparatus is simply a device for making visible the position of each one of these ions, so that the paths of individual cosmic-ray particles may be traced and photographed. If the gas or air through which the particle passes is saturated with water or alcohol vapor and is enclosed in a vessel which can be suddenly expanded by the proper amount just after the passage of the particle, a small droplet of water or alcohol will condense on each one of the ions present. Hence there appears in the wake of the particle a string of droplets which in a strong light will have the

appearance of a string of miniature beads and can be readily photographed. So although we cannot actually see a cosmic-ray particle we can do the next best thing—we can see clearly the exact path it has traversed in its passage through the gas, and we can obtain a permanent record of this path on a photographic film.

Several hundred thousand such cloud-chamber photographs of cosmic-ray particles have been made during the past ten years

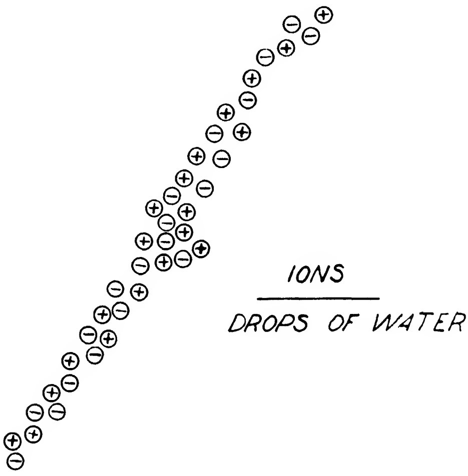


FIG. 62. Diagram indicating positive and negative ions formed in the wake of a cosmic-ray particle as it passes through a gas. By means of a cloud chamber a droplet of water is produced around each ion, thus making it possible to photograph the paths of cosmic ray particles.

in various laboratories both in this country and abroad. Like the electroscope the cloud chamber has been developed and improved until today its operation is wholly automatic. A cosmic-ray electron when it enters the cloud chamber makes its presence known by means of a device known as a Geiger counter to be described below. These Geiger counters, when placed one above and one below the cloud chamber, can be employed to control its operation. The passage of a single electron through the cloud chamber

is then sufficient to "trigger-off" the apparatus, the chamber is expanded, the source of illumination turned on, the camera snapped, etc., all automatically, and this occurs only when a particle enters the chamber with such a position and such a direction as to make a suitable photograph. Cloud-chamber photographs, several examples of which we shall see later, have given the physicist the most direct picture of the events which accompany the passage of cosmic-ray particles through matter.

The Geiger Counter

Geiger counters form the third general kind of experimental technique which has been successfully applied to cosmic-ray studies. The Geiger counter in its usual form consists of a metal cylinder along the axis of which is stretched a fine wire. The space within the metal cylinder contains air or some other gas at reduced pressure. When the tube is suitably prepared, and the proper voltage applied between the cylinder and central wire, a rapid discharge of electricity occurs between wire and cylinder upon the passage through it of a cosmic-ray particle. After suitable amplification these discharges can be employed to operate relays or other devices as is done when the Geiger counters are used to control cloud chambers. In other experiments recording devices are connected to the Geiger counters to record the number of cosmic-ray particles and hence the cosmic-ray intensity. Several Geiger counters may be connected together and arranged so that particles coming from only certain specified directions will be recorded. Such devices are the so-called cosmic-ray "telescopes." Experiments using Geiger counters have been made on mountain tops, in balloons, in deep mines, and are especially useful when information concerning the directional properties of the rays is desired. Modern Geiger counters, like the electroscopes and cloud chambers, are completely automatic in operation and yield records of their measurements on photographic films.

All three of these general techniques used in cosmic-ray researches were first developed and used in other fields of physical

research and later adapted for cosmic-ray experiments. The electroscope alone was used for cosmic-ray experiments up until the year 1927, when both the cloud chamber and the Geiger counter were first employed.

THE DISCOVERY OF COSMIC RAYS

The very early experiments on cosmic rays were performed before it was known that the effects observed were not caused by a radiation originating from a source close at hand, as for example from some material in the immediate vicinity of the instrument. The rays which emanate from radium and other radioactive substances ionize the molecules of a substance through which they pass and therefore can be recorded by electroscopes, cloud chambers, and Geiger counters. About the year 1900 when the study of radioactivity was in its infancy many investigators were using electroscopes to study the radiation emitted by radioactive materials. All investigators found, however, that their electroscopes continued to discharge, although at a very slow rate, even if their instruments were isolated from all known sources of radiations arising from radioactive materials. This in itself was not surprising for it was soon found that all materials of whatever origin contain a minute, but nevertheless detectable, amount of radium or other radioactive substance, even the materials of which the electroscope themselves were constructed. But some of these investigators, in studying this so-called "natural leak" of electroscopes, surrounded them by tons of lead in efforts to screen off as much as possible the rays from the radioactive materials which are present in the stone walls of the laboratory or even floating around in the atmosphere in minute quantities. In other experiments electroscopes were carried far out on the surface of frozen lakes in order to take advantage of the screening effect of the ice and water on the rays emanating from radioactive materials present in the earth.

The experiments all indicated the presence of another radiation, extremely weak in its intensity, in addition to that which

could be traced to radioactive materials. In 1909 Father Wulff, a Jesuit priest, on the trail of this effect, carried an electroscope to the top of the Eiffel Tower in Paris and there found almost the same rate of discharge as he found much nearer the earth's surface. In 1910, Gockel, a Swiss physicist, carrying an electroscope ascended nearly three miles in a balloon and found indications of an actual increase in intensity at the greater heights. This experimental result was indeed surprising and almost unbelievable in terms of the experience and knowledge of that day, but subsequent experiments in Germany by Hess and a few months later by Kolhorster, in which larger balloons were used and greater heights reached, proved without a doubt the reality of the increased effect at high altitudes. So by the time of the outbreak of the first World War the existence of this new type of radiation had been well established. It does not arise from any substance on the earth itself, but comes in from outside the earth and is partly absorbed in passing through the earth's atmosphere. At that time practically nothing was known of the properties of these new rays except that they possessed a far greater power of penetrating material substances than any ray of earthly origin, including even the most highly penetrating, the gamma rays arising in radioactive substances. These new rays later came to be known as cosmic rays.

During the period of the first World War research on cosmic rays, in common with most other scientific activity, suffered a severe decline, and not until 1922 was the next important advance made when a balloon was sent high up into the stratosphere. This flight showed the strength of the cosmic rays to be much greater in the stratosphere than at lower altitudes. Then began a renaissance of cosmic-ray research which has continued with increasing vigor up to the present time. This more recent period in the study of these baffling rays has to its credit a long series of accomplishments and has attracted the attention of scores of new investigators who have joined in their study in all parts of the world. Space will not permit here a detailed discussion of these

new experiments, nor even mentioning by name those who have made the largest contributions to this new body of knowledge.

The remainder of the paper will be devoted to a brief outline of some of the important characteristics of the cosmic rays that have been brought to light by the more recent experiments. First it will be shown, in a general way, how the recorded intensity of the radiation varies as the instruments are carried to greater and greater heights above sea level, and then how the intensity changes as the instruments are carried over different regions of the earth's surface. And lastly, we will take a peek at some of the events which transpire in the submicroscopic world of the elementary particles of matter as they are acted upon by the incoming cosmic-ray particles.

Variation with Altitude and Latitude

The many journeys of electroscopes into the high mountains and into the stratosphere have shown that the cosmic-ray intensity increases rapidly as one proceeds to greater and greater heights

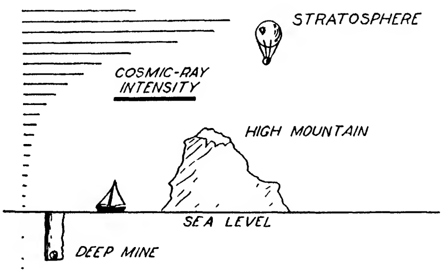


FIG. 63. Illustrating the intensity of cosmic rays at different levels.

in the atmosphere, until at an elevation of 80,000 feet it reaches a maximum value some 200 times greater than at sea level. Similarly as one goes down below the surface of lakes or into deep mines the intensity decreases rapidly until at depths of a few hundred feet below the surface of the earth the intensity

is found to be only one part in 10,000 of its value at sea level. (Fig. 63.)

The general decrease in the intensity from its maximum value in the stratosphere to its very small values below the surface

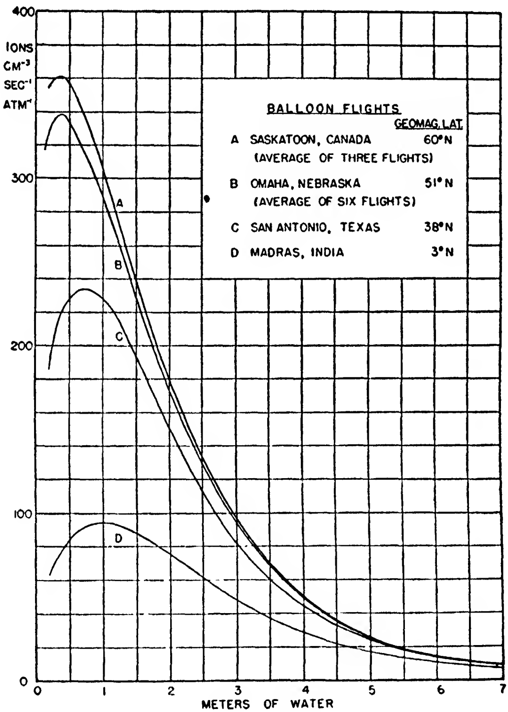


FIG. 64. Results of balloon flights at four different latitudes, showing the increase in the intensity of cosmic rays with altitude. (Courtesy Bowen, Millikan, and Neher.)

of the ground is a result of the fact that the rays interact with all matter through which they pass and are thus gradually absorbed. These experiments have shown that the cosmic rays are far more penetrating than any other known type of ray. The study of the

details of these processes of absorption of the cosmic rays by matter is one of the chief means of gaining information about the properties of the rays.

The increase in cosmic-ray intensity with increasing altitude at four different latitudes is shown in Figure 64. The altitude is here expressed in terms of the height in meters of a water column which would exert a pressure equal to the atmospheric pressure at

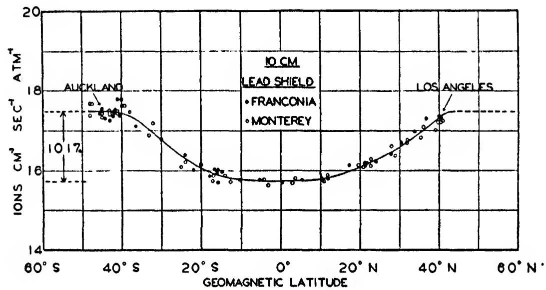


FIG. 65. Plots of measurements obtained from two trips between the United States and New Zealand, showing the "equatorial dip" of cosmic-ray intensities at sea level. (Courtesy Millikan and Neher.)

that altitude; sea level corresponds to about ten meters of water equivalent.

The round-the-world trips of the electroscopes on shipboard have revealed another interesting feature of the radiation. All the experiments show that as the ship approaches the equatorial regions of the earth there is a marked decrease in the cosmic-ray intensity. If the ship proceeds across the equator and into the southern latitudes the intensity then again rises until it reaches the same value it has in the northern latitudes. This change in intensity which amounts to about ten per cent if the measurements are made at sea level is indicated in Figure 65. If corresponding measurements are made at high altitudes the change with latitude is markedly greater—the strength of cosmic rays at a height of fifty thousand feet above sea level near the equator is less than

half the strength at the same height in the northern or southern regions of the earth. This change in intensity with latitude is the result of the fact that the earth itself is a great magnet. The incoming cosmic-ray particles, most of which are electrically charged, are deflected from their straight-line path by the magnetic force which comes into play when they pass into the earth's magnetic field. This magnetic field extends out into space several thousand miles above the surface of the earth. Many of the incoming particles are deflected into curved paths but are still able to reach the earth. Others are deviated to such an extent that traveling in great circular arcs they approach to within several hundred miles of the earth's surface, but miss striking it altogether, and then continue their journey through space in a new direction.

Studies of this effect of the earth's magnetic field on the cosmic rays have served to measure the energies of the individual high-speed particles which constitute the cosmic rays and to show that the bulk of the cosmic rays which impinge upon the top of the atmosphere consist of particles electrical in character. The individual particle energy is the most important of all the properties of the rays from the physicist's point of view. Also the magnetic field of the earth itself can be studied in these experiments. Measurements of the earth's magnetic field in regions hundreds of miles above the surface of the earth have thus become possible. An extension of these experiments promises to yield important information concerning the origin of the earth's magnetism, a subject which at present is still obscure.

In addition to these large-scale, world-wide studies of the cosmic rays, about which electroscopes and Geiger counters have furnished the most information, there are the equally interesting studies of those events which take place in the submicroscopic world where the elementary particles of matter play the important roles. For this phase of the work the cloud chamber has proved admirably suited.

The Interaction of Cosmic Rays with Matter

When cloud-chamber studies of the rays were begun in 1927 the existence of only two kinds of elementary particles was known—the electron of negative electric charge and the proton which bears a positive electric charge. At that time it was thought that all forms of matter were built up of various combinations of these two fundamental forms of matter. In 1932, however, a third

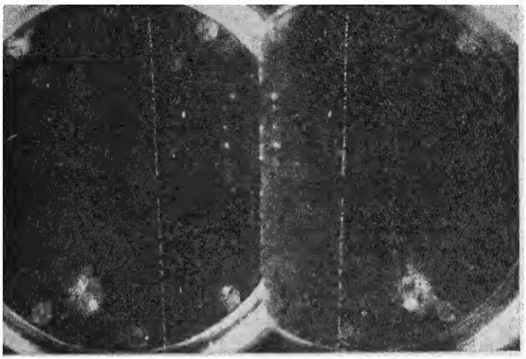


FIG. 66. A typical cloud-chamber photograph of a cosmic-ray particle. The left-hand image is the direct view; the right-hand image is made after reflection from a mirror and is employed in stereoscopic observation. The nearly vertical white line, showing a beaded structure, is the track of the cosmic-ray particle. Cf. Fig. 62.

member was added to this group when the neutron was discovered. Later that same year the discovery of the positron, or electron of positive electric charge, added a fourth member to this exclusive group of fundamental particles. This latter discovery came directly from cosmic-ray experiments, and we shall presently see what an important part the positron has in the phenomena which result from the passage of cosmic rays through matter. Within the space at my disposal it is not possible to give a detailed and complete discussion of all the phenomena which accompany the absorption of cosmic rays. Furthermore, the whole story could not in any case be told at the present time for there still

remains a great deal for the physicist to learn about cosmic rays.

We can, however, gain some insight into these occurrences by an examination of a selected group of several cloud-chamber photographs.

Figure 66 shows a typical cloud-chamber photograph of a cosmic-ray particle. Two images of the photograph are shown; the left-hand one is a direct view of the cloud chamber, the right-hand one was made after reflecting the light from a mirror and therefore appears reversed. By means of two such images it is possible to view the photographs stereoscopically and so perceive the tracks in a space of three dimensions. The nearly vertical, beaded white line is the string of water droplets condensed around the ions left in the particle's wake. Thus it shows the exact path followed by the particle in its passage across the chamber. The scale of the photographs is given by the circular outline which marks the walls of the cloud chamber whose diameter is six inches.

FIG. 67. A photograph similar to Fig. 66 but showing the curved path followed by the cosmic-ray particle when it traverses a strong magnetic field.

A knowledge of the energy or the speed of the particles is of great importance. This can be conveniently achieved by placing the cloud chamber between the pole faces of a powerful electromagnet. The particle will then not travel in a straight line but will follow an arc of a circle, and the greater the speed or energy of the particle the greater will be the radius of the circular arc. Figure 67 is a photograph taken in a magnetic field in which the curvature of the particle's path can be readily seen and measured. Measurements of this type have shown that the speeds, or energies, of the cosmic-ray particles are very great indeed, ranging up to values more than a thousand times as great as those of

any other particles known to the physicist. The enormously high energies of the cosmic-ray particles are perhaps their most striking characteristic. The effects of particles such as these when they pass through matter are largely dependent upon their energy, and therefore several important and highly novel discoveries have already resulted from studies of cosmic-ray particles.

A common technique in the experiments has been the placing

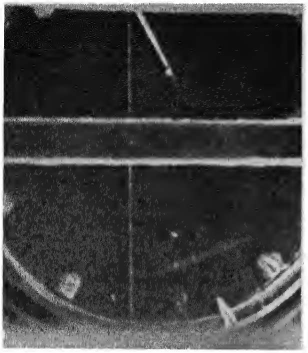


FIG. 68. A cosmic-ray particle passes downward through a plate of lead about one-half inch in thickness.

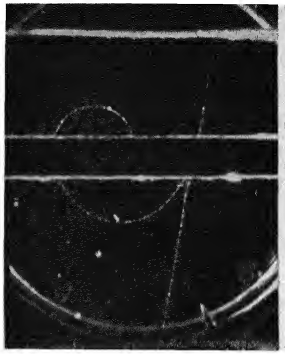


FIG. 69. The track of a cosmic-ray particle passing through a plate of carbon and striking an electron in the carbon. The struck electron produces the sharply curved circular track shown below and above the carbon plates.

of a barrier in the path of the particles. The barrier, for example, may take the form of a plate of lead of about one-half inch in thickness. Figure 68 shows that a particle may readily pass through such a plate with apparently no measurable diminution in its speed. Other experiments have shown that cosmic-ray particles are able to pass through solid lead more than a yard thick.

Occasionally, however, when a particle passes through such a plate other subsidiary phenomena are observed. In Figure 69,

for example, an electron of an atom in the plate was struck by the high-speed cosmic-ray particle as it passed through the plate, much as one billiard ball might be struck by another. The struck electron of the plate did not attain as high a speed as the incoming particle, as can be seen from the high degree of curvature which it shows in the magnetic field. It did, however, attain sufficient speed to enable it to emerge from the plate, to pass

FIG. 70. A shower of positive and negative cosmic-ray electrons. Cf. Fig. 71.

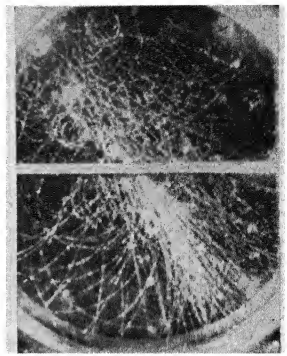


FIG. 71. A large shower of more than one hundred positive and negative electrons.

through it in the opposite direction, and then emerge again from its upper side, but with a further reduction in speed as indicated by the increased curvature. The plate in this case consisted of carbon, a relatively light material; if it had been lead the particle would not have emerged but would have been brought to rest somewhere within the plate. Continued studies of these so-called secondary electrons are yielding results of considerable importance.

The next photograph, Figure 70, shows a happening of still greater complexity. Here the paths of several electrons are indicated. It is evident that some of the electrons are positively charged and others negatively charged, as shown by the opposite

directions they are deflected in the magnetic field. This phenomenon in which several particles are observed at one time is called an "electron shower" and results from the births of many positive-negative electron pairs. Electron showers sometimes rise to the magnitude of veritable "cloudbursts" as shown in Figure 71 where the tracks of more than one hundred positive and negative electrons may be observed. This last photograph was made

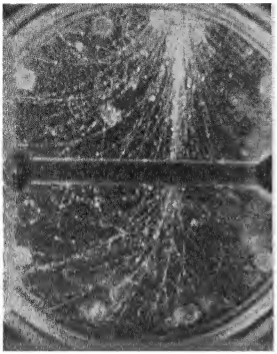


FIG. 72. A shower of electrons showing production of additional shower particles in a plate of platinum.

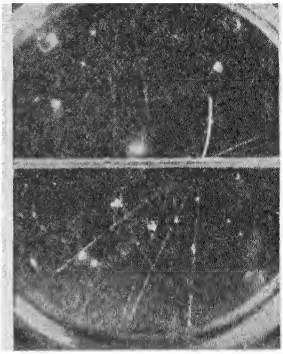


FIG. 73. The disruption of a lead atom by cosmic rays. The six tracks radiating from a point in the lead plate are produced by the products of the disintegration.

on the summit of Pike's Peak, 14,000 feet above sea level, where electron showers occur much more frequently than they do at sea level. Figure 72 shows another example of a large electron shower. In this experiment a plate of platinum, about one-half inch thick, was placed across the chamber. It can be seen that subsidiary showers arise in the platinum plate as a result of the action of the electrons which strike its upper surface.

Figure 73 shows another phenomenon which at the present time is not completely understood. Here the evidence points clearly to the fact that an atom of lead within the plate has been

disrupted by the action of the cosmic rays. Atomic disintegrations of this type differ from those usually produced by high-speed particles produced in the laboratory, as, for example, by means of a cyclotron, in that in the disintegrations by cosmic rays several particles, instead of the usual two, are thrown out in all directions, each with a high energy. In this photograph six particles are seen to have emerged from one point in the plate of lead. So far it has not been possible to determine with certainty the identity of all the products of the disintegrations produced by cosmic rays. This field will provide important material for future investigations.

In Figure 74 instead of a plate a Geiger counter was placed within the cloud chamber. Here one sees a remarkable and rare event. A particle comes in from above, passes through the Geiger counter and emerges with just enough energy to continue to travel through the gas of the chamber for about an inch; then it actually comes to rest in the gas of the chamber itself. The track which appears in this photograph has characteristics which distinguish it from the tracks produced by electrons or for that matter by any other known elementary particles. Here is cosmic-ray evidence which indicates that the physicist's list of elementary particles which now includes positive and negative electrons, protons and neutrons, may not be complete. Fundamental questions involving the identification of all the elementary particles of matter present in cosmic rays are now the subject of intensive investigations in several laboratories throughout the world.

THE POSITRON

A few words concerning the discovery of the positron and some of its properties may be of interest. The discussion will serve also to show the close relationship which exists between the results of cosmic-ray research and those researches in which radioactive materials are studied, or in which modern instruments, such as the cyclotron, are used to produce atomic disintegrations.

In 1932, shortly after the cloud chamber and magnetic field apparatus was first used for cosmic-ray studies, a series of several

thousand photographs was made in order to study the passage of cosmic-ray particles through matter. For this purpose a lead plate was placed in the chamber to act as a barrier for the particles, as shown, for example, in Figures 68 and 69. As stated above a particle in passing through this plate loses a part of its energy and emerges from the plate with a lower velocity than it possessed upon entering; hence the degree of the curvature in the magnetic field shows a difference, depending upon the amount of energy lost in the plate. Measurements made on the track of a particle before and after it has passed through a plate, together with observations of the density of the track itself, give definite information about the mass of the particle and the magnitude of the electric charge it carries.

Several photographs in this series clearly showed particles of positive electronic charge passing through the plate and emerging with lower energy. Measurements showed that none of these particles could be as massive as a proton, the only positive particle known at that time with a charge equal in magnitude to that of an electron. The photographs of these positively charged particles could be understood only if the particles were assumed to have a mass approximately equal to that of the ordinary electron of negative electric charge, and thus the first evidence for the existence of positive electrons, or positrons, was obtained.

As a result of experiments carried out in this country and in European laboratories it was soon discovered that it is not necessary to rely upon cosmic rays to produce positrons but that rays of laboratory origin can produce them as well. They have been found to occur when alpha rays from a radioactive substance are allowed to strike beryllium.[1] It has also been found that the well-known gamma rays from thorium can give rise to positrons. Determinations have been made of the relative numbers of positive and negative electrons produced by gamma rays from thorium, as well as of the energies with which electrons of both signs are ejected.

1. See E. O. Lawrence, "Atoms, New and Old," *Science in Progress*, 1939.

Here an observation was made which is of importance in deciding just how the positrons are produced. The most energetic positrons were found to have less energy than the most energetic negative electrons, by a definite amount corresponding approximately to 1,000,000 electron volts. The electron volt is a convenient unit for the measurement of electron energies and is equivalent to 1.6×10^{-12} ergs or 1.2×10^{-19} foot-pounds.

To point out the significance of the energy difference of 1,000,000 electron volts, it will be necessary to go back a few years and consider what the mathematical physicists had been thinking and doing while the other class of physicists, the experimenters, were performing their tests. One of the chief mathematical problems of the day was to incorporate the theory of relativity into quantum mechanics. The studies of Dirac in this connection resulted in his famous equation now known as the "Dirac electron equation."

This equation proved remarkably successful in solving a variety of problems which had hitherto baffled the theorists, but it contained one very striking feature which was a source of considerable annoyance. It required that electrons should under certain conditions have a negative energy and negative mass; they should have less than zero energy and weigh less than nothing. Dirac considered each point in space, including empty space or a perfect vacuum, to be "filled" with an infinity of such negative energy electrons. He also made the assumption that these negative mass electrons were unobservable and that it was a property of free space that they should be there. Dirac stated in 1930 that if one of these electrons should be removed, the "hole" in space that remained would manifest itself as an electron of positive electrical charge and of positive mass and energy.

The logic is perfect, for taking away less than nothing from space is equivalent to putting something there.

Because no positive electrons had ever been observed and because of a natural repugnance toward the idea that an infinity of electrons of negative mass should occupy each point in space,

practically all theoretical physicists considered this feature of Dirac's equation an unfortunate weakness. Because of the success of his equation, however, they continued to use it. But the discovery of the positron seemed to provide just the particle to correspond to one of these "holes" in space. The correspondence is indeed very close as is shown by the fact that in agreement with the Dirac theory, the fastest negative electrons had energies 1,000,000 electron volts greater than the fastest positive electrons.

This observation provides evidence for the correctness of the view that the positive and negative electrons might be created in pairs out of the incident radiation. Fortunately for the physicists of today, the theory of relativity as developed by Einstein shows that there exists a very close relationship between mass and energy—so close in fact that they may be considered as two aspects of the same entity. According to this view, if a pair of electrons is created out of the gamma radiation, then an amount of energy will be used up in the act of creation depending upon the masses of the particles formed. If one takes the known mass of the negative electron, and assumes the same value for the mass of the positron, then the calculated energy required for the creation of a pair of electrons comes out very close to 1,000,000 electron volts. The agreement with experiment is good, for according to this picture a negative electron can receive practically all the energy of the incident radiation; whereas a positron can appear only through the creation of a pair, hence the maximum energy it can receive is the energy of the incident radiation diminished by the 1,000,000 electron volts required to create the pair.

This bold theory of Dirac requires further that a positron, when it finds itself in a very ordinary environment, as, for example, in passing through water, shall, on the average, have only a very short life, of the order of a billionth of a second or less; for when a positron meets a negative electron, both particles will suffer the fate of complete annihilation, and in their stead a pair

of corpuscles of radiant energy, each of one-half million electron volts, will remain. Although the lifetime of positrons has not been actually measured, it has been shown to be very short, and the "annihilation radiation" announcing their death has been observed. The annihilation radiation is of the proper intensity and the energy of its individual corpuscles is approximately the required amount of one-half million electron volts, corresponding to the complete annihilation of the positrons. There is no reason to believe, however, that a positron, if removed from a region densely populated by negative electrons, may not live hundreds of millions of years, instead of perishing in a billionth of a second.

The creation of positive-negative electron pairs through the phenomenon described above is an important factor in the absorption of cosmic rays by matter. Because of the enormous particle energies which occur in cosmic rays the production of many electron pairs may result from the passage of a single cosmic-ray particle through the atmosphere or other absorbing material. Examples of the occurrence of large numbers of pairs, or "electron showers" as the phenomenon is now called, are shown in Figures 71 and 72. Experiments carried out in France and Switzerland have shown the occurrence of cosmic-ray showers, each consisting of many thousands of particles and covering an area 300 feet across. The energy required to produce such an enormous shower, all of which must have resided in a single particle reaching the earth's atmosphere from outside, is estimated to be of the order of 10^{16} electron volts. Such a prodigious concentration of kinetic energy is more than one hundred million times greater than the energies revealed by any other particles or rays except those of the cosmic rays. The difficulty in understanding the sources of energies of this magnitude is obvious, and accounts in part for the fact that the question of the origin of cosmic rays still remains a highly disputed and controversial matter.

So far as the interaction of cosmic rays with our own atmosphere or with other terrestrial materials is concerned the phenomenon of shower productions accounts fairly well for the manner in

which the less penetrating components of the cosmic rays are absorbed. The less penetrating or "absorbable" components of the cosmic-ray beam are therefore believed to consist largely of positive and negative electrons and photons which are absorbed principally through pair production giving rise to showers.

THE MESOTRON

Were it not for the presence also of other more highly penetrating particles in the cosmic rays the whole problem would be very much simpler. The experiments of the past several years have shown directly the presence of particles capable of passing through several feet of lead, and, indeed, other experiments have provided very strong evidence that certain particles are capable of passing through several hundred feet of a solid material such as rock.

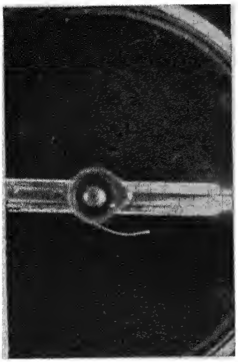


FIG. 74. The track of a cosmic-ray particle passing through a Geiger counter (center) placed in a cloud chamber, and coming to rest in the gas of the cloud chamber, as described on page 189. The degree of ionization, as shown by the heaviness of the track, indicates that this particle is a mesotron.

In the course of the development of cosmic-ray researches these highly penetrating components of the rays were ascribed first to photons and later to electrons. As the experimental and theoretical investigations of electron pair and shower production as discussed above were extended, strong evidence was brought forth that the processes of shower formation would place such severe limitations on the ranges of both electrons and photons of all energies; that they could not be used to account for the highly penetrating particles. Hence it was suggested that the penetrating particles were protons, but further investigations revealed difficulties with this point of view. The experiments and the arguments concerning the nature of the

highly penetrating cosmic-ray particles are too involved to be presented here, but suffice it to say that the only satisfactory explanation seems to be in terms of a new kind of particle, hitherto unknown to physics. This particle has been called a mesotron. It occurs with either a positive and negative charge equal in magnitude to that of a normal electron, but its mass seems to lie between that of the electron and the proton. Other kinds of data such as that shown in Figure 74 also provide strong evidence for the existence of particles of this type. Although the precise nature and many of the properties of this particle are unknown at the present time, there is universal agreement among cosmic-ray investigators that the hypothesis of the existence of mesotrons is required to explain the cosmic-ray observations.

Experiments indicate that cosmic rays at sea level consist almost wholly of mesotrons, most of which are produced in the upper atmosphere. The manner in which they are produced is completely obscure. Other experiments seem to show that the mesotron is short lived, disintegrating in about one millionth of a second presumably into an electron and a neutrino. Should future investigations provide positive proof of their disintegration, then in addition to the positron another transitory particle will have been added to the list of the elementary particles of physics.

CONCLUSION

The subject of cosmic rays represents in itself an extensive field of study, and at the present time is developing rapidly. It is yet too early to make a final appraisal of the value of cosmic-ray research to science as a whole, or to state the contribution which such studies will make to the other broad fields of science as, for example, to biology, meteorology, geology, or astrophysics. Experimental attempts have been made to detect a possible influence of cosmic rays on living organisms but these are entirely preliminary in character. Relationships between cosmic-ray effects and the state of the atmosphere have been established, and recently changes in cosmic-ray intensities have been cor-

related with the presence of fronts separating different air masses in the atmosphere. A continuation of the studies of the latitude effect will undoubtedly reveal new properties of the earth's magnetism and may assist in bringing about a more complete understanding of the phenomena responsible for the earth's magnetism. Evidence has been obtained of a possible influence of the magnetic field of the sun on cosmic-ray intensities, and an extension of these experiments may add to our present meager knowledge of the sun's magnetism. But it is in the great field of astrophysics that research on the cosmic rays gives the richest promise of contributing new knowledge. Just how important the contribution will be will be known only after cosmic-ray research has developed sufficiently to provide a clear understanding of the now mysterious processes which give rise to the rays. The most important results of cosmic-ray investigation so far have been in the field of physics, in particular, with regard to the new facts brought to light concerning the properties and the behavior of the elementary particles of matter.

VIII

THE MOTIONS OF IONS AND PROTEINS IN ELECTRIC FIELDS

By Duncan A. MacInnes

The Rockefeller Institute for Medical Research

The purpose of this discussion is to outline the present status of the ionic theory of solutions of salts, acids, and bases, including an interesting and important type of ions, namely, proteins. Also, although many forms of experiment have been used in obtaining the data on which the present theories are based, I wish to discuss particularly one kind of experimental technique, that of the *moving boundary* between two conducting solutions. It seems worth while to consider the material in this form in order to emphasize the fact that advances in science are largely due to the improvement of the tools of research. Thus to take only one instance, Laue's prediction, shortly verified by Friedrich and Knipping, that X rays would be refracted by passing through a crystal, has led, through the development of X-ray technique, on the one hand, to the whole modern advance of crystallography, and on the other to the study of the X rays themselves, and, indeed, to a large part of modern physics. I hope to show that in a similar way the moving-boundary method has given convincing evidence of the presence of ions in solutions; it has yielded data which have made it possible to verify, with accuracy, the more recent theories concerning ionization; and, in addition, has provided a method for studying proteins, and similar substances, that bids fair to be one of the most important means of biological and medical research.

This paper will, therefore, discuss subjects apparently as unrelated as the effect of fever on the composition of blood and the interionic attraction theory of electrolytes. The unifying idea is the fact that one type of experiment can yield results of interest and importance in as diverse fields as those just mentioned.

EVIDENCE FROM MOVING BOUNDARY MEASUREMENTS ON THE INTERIONIC ATTRACTION THEORY OF STRONG ELECTROLYTES

A few words of historical background may be helpful. In 1883 Arrhenius advanced the theory that in solutions of salts, acids, and bases the molecules are, largely at least, split up into positively and negatively charged *ions*. Thus, for example, an aqueous solution of sodium chloride has been conceived to consist, in large part, of positively charged sodium ions and negatively charged chloride ions, which may be represented by Na^+ and Cl^-. On its appearance the theory encountered vigorous opposition, based mostly on the apparently quite reasonable objections that sodium, which is well known to be a soft inflammable metal, and chlorine, which makes itself evident as an evil-smelling gas, can hardly be present in a solution of common salt. Also, if positive and negative ions are as close together as they must be in a solution they would certainly be expected to attract each other and recombine. In spite of these objections and others, the theory rapidly gained adherents since it accounted more or less adequately for the main facts about solutions of substances we now term electrolytes. The most important facts about such solutions are that they are conductors of electricity and, in addition, that they exhibit abnormal thermodynamic properties, which means that they produce exceptionally large depressions of the freezing points and vapor pressures of the solvents in which they are dissolved.

The mental picture with which Arrhenius worked was somewhat as shown in Figure 75. The symbols + and — represent the positive and negative ions, respectively. A part of the ions are

considered to be free, and another portion of the ions are in contact with each other, forming neutral molecules. The former were considered to conduct electricity, and the latter to have no influence on the conductance. The positions of the ions and neutral molecules would, of course, be rapidly shifting, due to thermal vibrations. Neutral molecules were thought to be constantly breaking up into ions, and the ions, in turn, uniting to form molecules, the proportion of ions, however, increasing as the solution was made more and more dilute. A salt solution was thus considered, by Arrhenius, to consist largely of free positive and negative ions which, under the influence of an impressed electric force, drift to the cathode and anode, respectively.

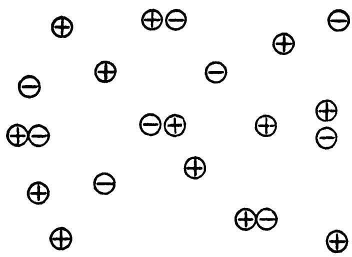


FIG. 75. Illustrating Arrhenius' theory of electrolytic dissociation.

Shortly after the appearance of Arrhenius' papers, Kohlrausch computed the velocities of the ions in a unit electric field, i. e., the mobilities, necessary to account for the observed conductances of salt solutions as actually measured. Some of Kohlrausch's figures are given in Table II.

TABLE I

Mobilities of Ions, in Centimeters per Second per Volt, at 18°C. According to Kohlrausch

Cations	H^+	K^+	NH_4^+	Li^+	$1/2\ Ba^{++}$	$1/2\ Mg^{++}$
	.00300	.00057	.00055	.00026	.00033	.00029
Anions	Cl^-	I^-	NO_3^-	ClO_3^-	$C_2H_3O_2^-$	
	.00059	.00060	.00053	.00046	.00029	

Thus, according to these figures, a potassium ion in a solution on which there is impressed a potential of 1 volt per centimeter should drift toward the cathode at a rate of about 2 centimeters

in an hour. However, these computations were theoretical. They accounted for the observed data on conductances of solutions and for the transference numbers that had been found, long before, by Hittorf, but experimental data showing that ions actually move at such velocities were lacking. It must be recalled that at the time the computations were made even the existence of the ions was in doubt. However, in 1886, Oliver Lodge [1] carried out some crude but very illuminating experiments, which tended to show that ions do actually move at rates of the order of magnitude indicated by Kohlrausch's computations. These were the first moving-boundary experiments, and they had much influence in

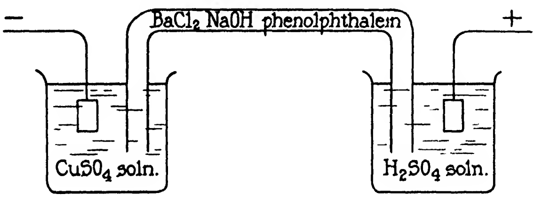


FIG. 76. Arrangement for Lodge's moving-boundary experiments.

the early acceptance of the ionic theory of electrolytes. One of Lodge's experiments is represented schematically in Figure 76. The anode and cathode vessels contained respectively copper sulphate and sulphuric acid. The tube connecting the two vessels contained barium chloride, set into a plant jelly (agar-agar), with the addition of a little sodium hydroxide and phenolphthalein, to give a red color. When the electric current was passed the red color bleached out in the tube at the two ends of the tube nearest the electrodes. Successive pictures of what happened to the color after the bleaching had progressed around the bends of the tube is shown in Figure 77, which is taken from Lodge's paper. It will be seen that one boundary moved from right to left, and may be regarded, very roughly, as indicating the advance of the hydrogen ion. Another boundary moved from left to right, and

1. Lodge, 1886.

though Lodge considered it to be due to the sulphate ion it is more probable that the change of color arose from the motion of the hydroxyl ion, especially as the rate of motion of the boundary is about half of that ascribed to the hydrogen ion. It is now known that hydrogen and hydroxyl ions have about those

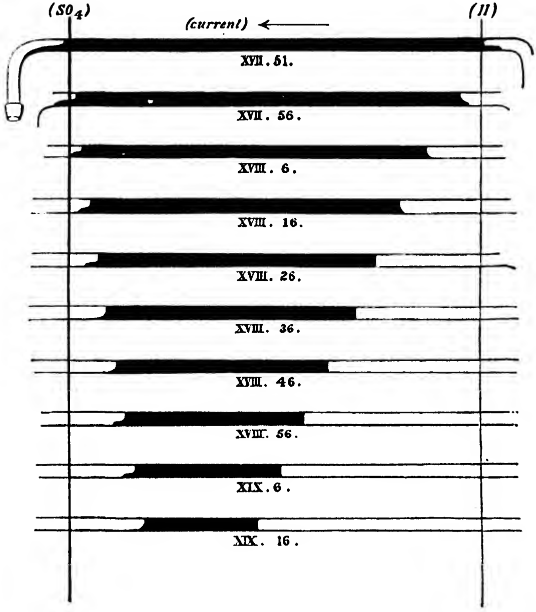


FIG. 77. Showing the motion of boundaries in Lodge's moving-boundary experiments.

relative mobilities. The important point was, however, that visual evidence was afforded by these experiments that boundaries, which could be ascribed to definite charged ions, moved with velocities of the order of magnitude predicted by Kohlrausch.

With this promising start, investigations of the motions of the ions were carried out by various investigators, including Whetham, Nernst, Denison and Steele, and Franklin and Cady. Along

with these experimental studies, theoretical papers appeared by Kohlrausch,[2] Weber,[3] Miller,[4] Laue,[5] and others. A number of years ago the author became convinced that precise values of ion mobilities would be of great help in testing the newer ideas connected with solutions of ions, and looking over the various available methods decided that the moving-boundary method might yield results of the desired accuracy. However, the experimental procedures, as published, proved difficult to work with. Two methods for forming the boundaries, which was the greatest difficulty, were finally developed, both of which worked smoothly. These are the "sheared-boundary" technique of MacInnes and Brighton [6] and the "autogenic boundary" of Franklin and Cady,[7] and of Cady and Longsworth.[8] The principle of the first of these is shown in Figure 78. The vessel *A* contains an electrode *E* and one of the two solutions to be brought into contact, and is arranged so that a pendant drop, *a*, emerges from the sliding plate *B*. The other solution is contained in the tube *D*, and another drop, *b*, emerges from the plate *C*. By bringing one of the tubes over the other, as shown in the lower part of Figure 78, the excess solution is sheared off, as shown at *e′* and *e* and a boundary is formed between the two solutions, very little disturbed by mixing or diffusion. Now if the two solutions have been correctly chosen and current is passed, the boundary will move down the graduated tube *D* and its speed can be accurately measured. The formation of an "autogenic" boundary may be illus-

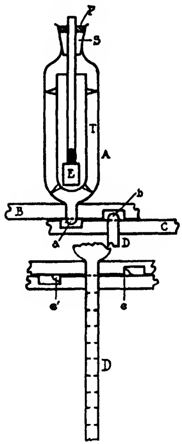


FIG. 78. The "shear" technique for forming moving boundaries as described on page 202.

2. Kohlrausch, 1897.
3. Weber, 1897, 1910.
4. Miller, 1909.
5. Laue, 1915.
6. MacInness and Brighton, 1925.
7. Franklin and Cady, 1904.
8. Cady and Longsworth, 1929.

trated by a specific case. Initially a tube containing, for instance, potassium chloride, is placed so that its lower end is covered with a disk of metallic cadmium, as shown in Figure 79, a. If now an electric current is passed in the direction indicated the metal will be attacked, forming cadmium chloride, and a boundary between the potassium and cadmium ions will pass up the tube, as indicated in Figure 79, b. These two simple devices for forming the boundaries have been largely responsible for making the determination of ion mobilities one of the most accurate measurements that can now be performed. It was early noted that it is not necessary to use colored solutions; the boundary may be followed by simple optical devices since the two solutions in contact will, in general, have different refractive indices.

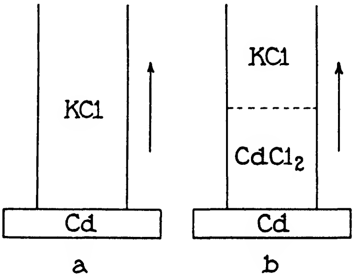


FIG. 79. The "autogenic" technique for moving boundaries as described on page 203.

It is a very fortunate circumstance that disturbing effects due to diffusion and convection are overcome by a "restoring effect," the nature of which may be made clear from the following example. Consider the case of a boundary between lithium chloride and potassium chloride, shown diagrammatically in Figure 80. Since the lithium ion has a lower mobility than the potassium ion and the lithium chloride solution is more dilute than the potassium chloride solution, the passage of current will cause a greater potential drop in the former solution than in the latter. This is also shown diagrammatically in the figure, where values of the electromotive force are plotted as ordinates and distances along the measuring tube as abscissae. Now if some of the relatively fast moving potassium ions diffuse or are carried by convection into the lithium chloride region, they will encounter a higher potential gradient and will be sent forward to the boundary. On

the other hand, if lithium ions diffuse into the potassium chloride region they will move slower than the potassium ions and will finally be overtaken by the moving boundary. After appreciable diffusion there will not, of course, be a sharp break in the potential lines, such as is shown in Figure 80. However, for this restoring effect to function it is only necessary for the gradient to be steeper behind a given solution layer than in front of it.

The question naturally arises as to why anyone should want to spend long years devising methods for making boundaries be-

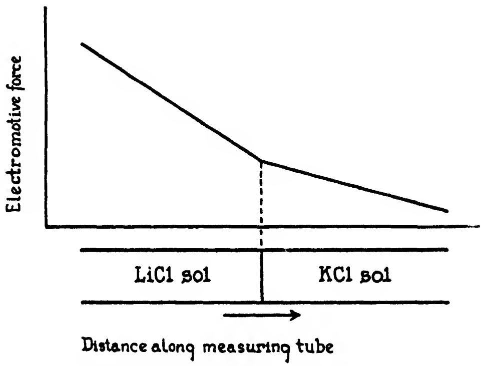


FIG. 80. The rectifying effect in moving boundaries between salt solutions.

tween salt solutions and then getting thousands of figures concerning the motions of such boundaries when a current is passed. I shall try and show just what its importance is. Instead of the speed or mobility of an ion, for many purposes it is more convenient to deal with the *transference number*. For an electrolyte which has only two kinds of ions the transference number, t, is the ratio of the mobility of one ion to the sum of the mobilities of both ions; thus the transference number t_K of the potassium ion in potassium chloride is given by the formula

$$t_K = \frac{U_K}{U_K + U_{Cl}} \qquad (1)$$

in which U_K and U_{Cl} are the mobilities of the potassium and chloride ions respectively.

It will be recalled that a great advance was made by Van't Hoff when he showed that a substance in solution has many properties analogous to those of a gas. Thus a solution of sugar has an osmotic pressure, with dilute solutions at least, not far from the pressure that would be measured if the same number of molecules were present, in the volume, as a gas. Though the idea had utility in connection with the understanding of many physiological processes there were quite a number of cases in which it did not fit. Thus electrolytes were found to give too large osmotic

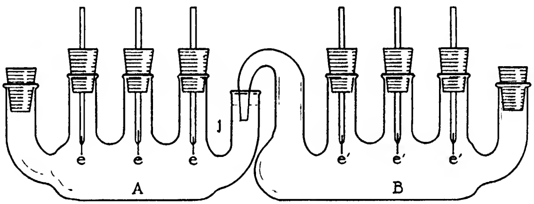


FIG. 81. A concentration cell.

effects. These observations and others led Arrhenius to his theory, already mentioned, of electrolytic dissociation. He found it useful to assume that a salt, such as potassium chloride, is dissociated into positive and negative ions in water solution and that the proportion of ions, α, called the "degree of dissociation," increases as the solution is more and more diluted. One way of studying the matter and getting at values of this degree of dissociation appeared to be to measure the potentials of concentration cells. A simple form of concentration cell is shown in Figure 81. Here vessel *A* contains the more concentrated of two solutions, of sodium chloride for example, and the vessel *B* the more dilute. In these solutions electrodes *e* and *e′* (in this case silver covered with silver chloride) are placed, several being used for comparison. A typical case of such a concentration cell can be conveniently represented by the symbols

$$\text{Ag; AgCl, NaCl} : \text{NaCl, AgCl; Ag} \qquad (2)$$
$$(C_A) \qquad (C_B)$$

Now if the ions are "ideal solutes" the electromotive force E between the electrodes in the two solutions should follow the equation

$$E = \frac{2tRT}{F} \log \frac{C_A \alpha_A}{C_B \alpha_B} \qquad (3)$$

in which t is the transference number of the positive ion, R is the gas constant, T the absolute temperature, and F the faraday, C_A and C_B are the salt concentrations in vessels A and B, respectively, of Figure 81, and α_A and α_B are the corresponding degrees of dissociation. It was for a generation an annoyance and puzzle to everybody in the field of physical chemistry that equation (3) could not be made to work if degrees of dissociation α obtained from conductance measurements, according to Arrhenius' theory, were used. This is one way of stating that there was an "anomaly of strong electrolytes" which meant that kinetic theories of electrolytic conductance and the ideas then current concerning the thermodynamic properties of the solutions could not be reconciled. A way of sidetracking this difficulty was suggested by G. N. Lewis. In effect he rewrote equation (3) as follows:

$$E = \frac{2tRT}{F} \log \frac{C_A f_A}{C_B f_B} \qquad (4)$$

The f values in this equation are called "activity coefficients." For a long time the f values were purely empirical. They were, however, useful since if once determined they could, with suitable precautions, be used in any relevant thermodynamic equation. For moderately concentrated solutions these activity coefficients are less than unity but approach unity as a limit as the solutions become more and more dilute.

Now the tacit assumption that Arrhenius, Ostwald, and their followers made was that the ions are perfect solutes, in spite of the fact that they were also assumed to be highly charged and, necessarily, very close to each other. However, first Milner [9] and later Debye and Hückel,[10] in epoch-making papers, showed theoretically at least, that the variation of the activity coefficients of the ion constituents from unity can be explained by the electrostatic attractions and repulsions of charges carried by the ions. Positive ions will attract negative ions, and like-charged ions will repel each other. Thus if these electrostatic forces alone were active the ions would arrange themselves in a regular "space lattice" as in a crystal. Due to the disturbing effect of the heat vibrations of the ions and of the solvent molecules, no regular arrangement of the ions can persist. If, however, you were taking a ride on one of the positive ions more of your neighbors would be negative ions than positive ones, and this would be more evident the greater the concentration. Thus in a salt solution the ions are not quite independent of each other, and their thermodynamic properties will therefore be expected to change with the concentration.

Using these simple assumptions and ingenious mathematical-physical reasoning, Debye and Hückel arrived at the formula

$$-\log f_i = \frac{A\sqrt{C}}{1 + Ba_i\sqrt{C}} \tag{5}$$

in which f_i is the ion activity coefficient, A is a constant computable from universal constants, B also contains such constants, and a_i is the size of the ions, or more precisely speaking, the "distance of closest approach" of the positive and negative ions. The constants, but not the form of the equation, change when the type of electrolyte is varied.

We have made at the Rockefeller Institute what we believe

9. Milner, 1912, 1913.

10. Debye and Hückel, 1923.

to be the most precise tests yet obtained of the Debye-Hückel theory of ionic interaction [11] for the region of concentration for which the underlying assumptions would be expected to hold. For not too highly charged ions, the theory fits with the experiments within the limits of a very small experimental error. As can be seen from equation (4) the transference number t is an important factor in obtaining activity coefficient ratios, f'/f'', for the test.[12] Precision transference numbers have been determined by Dr. L. G. Longsworth [13] by the moving-boundary method. These transference numbers are not, however, constants, as assumed in obtaining equation (4), but are functions of the concentration C, which somewhat complicates the computation. The actual equation used for solutes of the type of NaCl was, therefore,

$$E = \frac{2RT}{F} \int_A^B t \, d \log Cf \qquad (6)$$

in which allowance is made for this variation of the transference number, t.

Some of the results are represented in Figure 82, where the negative of the logarithm of the activity coefficient is plotted against the square root of the concentration. The distance of closest approach, a_i, computed from the data is seen to vary from 5.6 Ångström units for hydrochloric acid to 2.3 Ångström units for silver nitrate. It will be seen from the plot that the agreement of observed and computed values represented by the smooth curves is very close. As a matter of fact the agreement of the observed values of f_i and those computed from equation (5)

11. Brown and MacInnes, 1935; Shedlovsky and MacInnes, 1936, 1937, 1939.

12. Tests of the Debye-Hückel theory have also been made with concentration cells "without transference," i. e., that do not require transference numbers. Only in exceptional cases, however, can they be used with dilute enough solutions for the purpose.

13. Longsworth, 1935; Longsworth and MacInnes, 1938.

agree within a few hundredths of a per cent up to a concentration of about .05 normal. It is evident that, for the substances mentioned, the new theory accounts adequately for the experimental results, both for electrolytes of the 1-1 type of KCl and for the more complicated 2-1 type represented by calcium chloride. If, as assumed in Arrhenius' theory, the ions had no influence on each other the activity coefficient, f_1, would be unity. This would correspond to the straight line forming the lower border of the diagram. Figure 82 also contains the straight lines representing the "limiting laws"

$$-\log f = A\sqrt{C} \tag{7}$$

for the two types of electrolyte which hold for exceedingly dilute solutions. In obtaining equation (5) a very important simplifying assumption was made. Instead of postulating, as Arrhenius did, that the dissolved electrolyte is partly present as ions and partly as undissociated electrolyte, it was found, for strong electrolytes at least, that agreement between theory and observation could be more readily obtained if it were assumed that the electrolyte is *all* present as ions. To a generation of physical chemists who were brought up under the influence of Arrhenius, the idea that electrolytes could be completely dissociated seemed quite revolutionary, but by this time most of us have gotten used to it. With our modern knowledge of solutions we are presented with the opposite difficulty, that of accounting for many cases of incomplete dissociation. Thus an

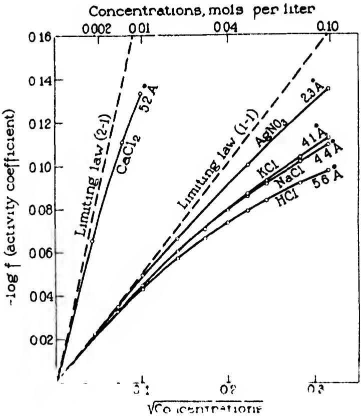


FIG. 82. Illustrating the adequacy of the Debye-Hückel theory for the dependence of activity coefficients on concentration.

electrolyte with ions the size of those of acetic acid should, according to the ideas of today, be totally ionized in water solution, whereas, as a matter of fact, it is but slightly ionized. This means that processes are going on in the molecule that we do not yet understand.

The success of the interionic attraction theory in the thermodynamic field has been accompanied by conquests in the region of the kinetic properties of solutions, particularly their conductivity. It will be recalled that the equivalent conductance, Λ, of a strong electrolyte, such as sodium chloride, decreases as the concentration is increased. Arrhenius and his followers considered this decrease of equivalent conductance to be due to a falling off in the number of carriers of electricity. A portion of the oppositely charged ions were supposed, in increasing degree as the concentration is raised, to stick together, as indicated in Figure 75, and thus take no part in the conductance. We have already seen that the present notion is that a solution of a strong electrolyte consists entirely of ions. There must, obviously be some other explanation of the observed decrease in conductance. Debye and Hückel, and Onsager, have, as a matter of fact, shown that there are two "effects" or mechanisms tending to bring about decreases in the equivalent conductance of an electrolyte with increase of concentration. Both of these arise from the attractions and repulsions of the charged ions. These are (*a*) the *electrophoretic effect* and (*b*) the *time of relaxation effect*.

One of the consequences of the interionic attraction theory is that each ion carries around with it a so-called "ion atmosphere" due to considerations that have already been discussed. If a positive ion has, on the average, more negative than positive ions near it the result will be the same as if a charge of negative electricity were spread out symmetrically in all directions in the solvent, most of it lying near the chosen ion but spreading out more and more thinly as the distance from the ion increases. Although in the mathematical discussions of this potential it is treated as a reality, it is actually the result of a time average of a distribution

of the ions, each ion serving as a center of an ionic atmosphere and also as a part of the ionic atmosphere of other ions. Now if a potential gradient is impressed on the solution, as in an electrolysis or a measurement of conductance, the charges of the ionic atmosphere will tend to move, carrying solvent with them, in a direction opposite to that of the chosen ion. This countercurrent of solvent will impose a drag on the ion, tending to decrease its velocity. This is known as the "electrophoretic effect." The second effect is possibly more difficult to visualize, but an attempt is made in Figure 83. If we move an ion, represented by the black circle, suddenly from position *a* to position *b* its ionic atmosphere will go along also in the same direction. However, it takes some time for a message that something has happened at the point where the ion is located to get to the outlying regions of the ion atmosphere. These unnotified regions (outside the dotted circle in the figure) are unsymmetrically spaced around the ion and thus exert a pull on the notified regions. This gives rise to an extra drag on the central ion. This picture gives a rough idea of the "time of relaxation effect."

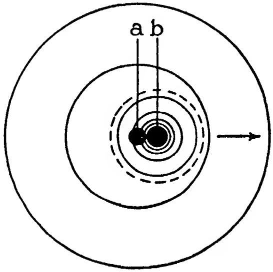


FIG. 83. Illustrating the "time of relaxation effect."

There is a very sensitive test that can be made of these newer conceptions, using transference numbers. According to the theory of Arrhenius, the ions, being quite independent of each other, should have the same mobilities at all concentrations, and, therefore, their transference numbers should remain constant. The interionic attraction theory, however, predicts that these transference numbers should vary and, furthermore, for very dilute solutions at least, how much they should change with the concentration. Figure 84 shows the results that have been obtained in this field by Dr. Longsworth at the Rockefeller Institute. Here the measured transference numbers are plotted against the

square root of the concentration, and the curved lines have been drawn through these points. The change of the transference number predicted from the interionic attraction theory is represented in each case by the straight line starting from the value of the limiting transference number, at zero concentration. It will be

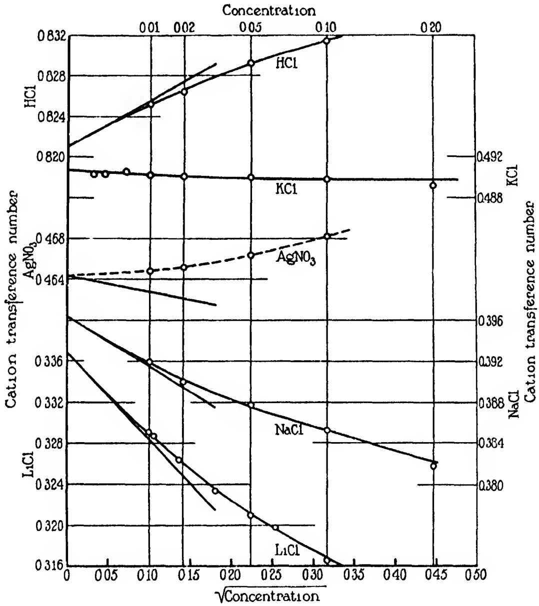


FIG. 84. The theoretical and observed variation of transference numbers with the concentration.

seen that the curves passing through the observed points merge naturally into the theoretical line, with the exception of that for silver nitrate which substance is, in several respects, abnormal.

Another use of the precision transference numbers made available to us by the moving-boundary method is in obtaining ion conductances of the ion constituents. The conductance of the chloride ion, λ_{Cl}, from sodium chloride for instance, is given by the relation

$$\lambda_{Cl} = t_{Cl}\, \Lambda_{NaCl} \tag{8}$$

in which t_{Cl} is the transference number of the chloride ion, and Λ_{NaCl} is the equivalent conductance of sodium chloride. The relation of the equivalent conductance of the chloride ion to its mobility, U_{Cl}, is given by the simple formula

$$U_{Cl} = F\lambda_{Cl} \tag{9}$$

in which F is the value of the faraday. The values of the equivalent conductance of the chloride ion from various chlorides at different concentrations are given in Table III. The data are based on the precise conductance measurements of Dr. Shedlovsky at the Rockefeller Institute laboratory. A number of interesting conclusions can be arrived at from a study of these

TABLE II

Equivalent Conductances of the Chloride Ion Constituent from Different Electrolytes at 25°

	Concentration, equivalents per liter						
Electrolyte	.001	.002	.005	.01	.02	.05	.10
KCl	74.88	74.28	73.22	72.07	70.56	68.03	65.79
NaCl				72.05	70.54	67.92	65.58
HCl				72.06	70.62	68.16	65.98
LiCl				72 02	70.52	67 96	65.49

figures. Although the values of the equivalent conductance differ somewhat at .1 normal, they agree almost within the experimental error at .01 normal. It would take us too far afield at this point but it can be shown that all the figures are in accord with a simple extension, due to Dr. Shedlovsky, of Onsager's [14] equation, which is derived from the interionic attraction theory, and that they all lead to a value of the limiting equivalent conductance of the chloride ion of 76.34 at 25°. These results, and many others of similar kind, tend to confirm the picture of solutions of strong electrolytes as consisting of independent ions, whose electric fields, however, interact in such manners as to

14. Onsager, 1926, 1927.

affect each other's thermodynamic activities and their mobilities in applied electric fields.

From this very brief outline I hope that I have shown that we have a fairly satisfactory theory for strong electrolytes. If a series of discussions, instead of a portion of one, could be devoted to the topic it would be possible to show that the interionic attraction theory is equally useful in dealing with the results of the measurements on weak electrolytes. Such measurements have been obtained at the Rockefeller Institute and at Yale by Harned and his associates. Particularly in the hands of Fuoss and Kraus, the theory has been of great service in interpreting the data on nonaqueous solutions. A more complete discussion of the interionic attraction theory should include the researches of La Mer and associates, on electrolytes of higher valence types and on the theoretical extensions of the Debye-Hückel theory. Such a discussion should also include the Wien and Debye-Falkenhagen effects, which deserve extended treatment. It should not, however, be concluded that no problems remain to be solved in connection with solutions of electrolytes. Much remains to be done, particularly for electrolytes involving highly charged ions, and for concentrated solutions.

EVIDENCE FROM MOVING BOUNDARY MEASUREMENTS AS TO THE PHYSICAL CHEMISTRY OF PROTEINS

We shall next take a jump to another field in which active research is now going on, but one where only a report of rapid progress instead of a record of final accomplishment can be given. The field I refer to is that of the physical chemistry of proteins. As frequently happens, the Greeks had the word for it. Protein means "to be first," which adequately expresses the importance of proteins in the substance of our bodies in health and disease and in our daily nourishment. A vast amount of research has been carried out on protein materials. It is well known that

they can be broken down into various amino acids in addition to prosthetic groups. Except that the amino acids are probably arranged into peptide chains we have really little knowledge as to their internal structure, though some good guesses have been made which are serving as bases for further research.

There are quite a number of ways in which the problem of the constitution and physical chemistry of proteins is being attacked. Some of these are as follows: the analysis of proteins into their constituents, represented by the work of Bergmann, Vickery, and others; the study of the arrangement of the atoms in crystalline proteins by the means of X rays by Bernal and his school in England; the use of titration data by Cannan; of solubility, dielectric constant measurements, pH measurements, etc., by Cohn and his associates; the ultracentrifuge by Svedberg; and many others that might be mentioned. The range of the attack on the problem is so great that the participants hardly speak each other's scientific language.

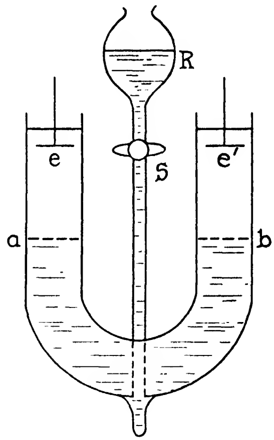


FIG. 85. Burton's apparatus for studying electrophoresis of colloidal solutions as described on page 216.

The method of study that I wish to discuss is, however, that of electrophoresis. The term electrophoresis (or cataphoresis) has long been used in connection with the motion of colloidal particles in electric fields. There are two ways of studying electrophoresis. The microscopic method may be used if the particles are large enough to be seen in a microscope. This method was first used by Ellis,[15] and later by Northrop and Kunitz,[16] and by Abramson.[17] The motion of colloidal particles, as well as of ions, may, however, also be followed by the moving-boundary method.

15. Ellis, 1912.
16. Northrop and Kunitz, 1924–25.
17. Abramson, 1928–29.

This was done as early as 1906 by Burton,[18] who used the simple apparatus shown diagrammatically in Figure 85. By first introducing the pure solvent into the U-tube, and then cautiously opening the stopcock (*S*) between the U-tube and the reservoir (*R*) containing the colloidal solution, the electrodes (*e-e′*) can be surrounded by the solvent, and two fairly sharp boundaries (*a* and *b*) between the solvent and the colloidal solution are formed. On applying sufficient potential these boundaries will move. A serious difficulty with early experiments with this method was that the boundaries were disturbed by convection currents arising from the heating effect of the electric current since, unlike boundaries between properly chosen salt solutions, there is no "restoring effect" that maintains the sharpness of the boundaries. Protein solutions were investigated by this method by early workers, but it was found difficult to follow the boundaries because such solutions are usually transparent.

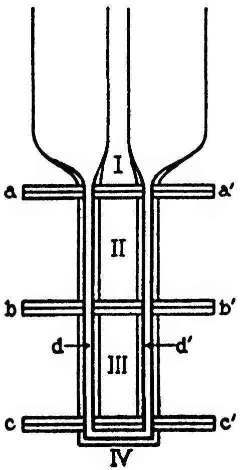


FIG. 86. Tiselius' cell for electrophoretic measurements as described on page 217.

Due to the investigations of Arne Tiselius,[19] however, the moving-boundary method for studying proteins and related materials has been developed into one of the most useful tools of physical chemistry and of biology. Tiselius' main additions to the method are: (*a*) the use of the "shear" technique in forming the initial boundary; (*b*) the adapting of the Toepler schlieren effect to make the boundaries visible; and (*c*) the minimizing of convection effects by the use of a low temperature thermostat. These developments will be discussed in the order given.

The type of cell devised by Tiselius for forming the boundaries

18. Burton, 1906. 19. Tiselius, 1937.

and providing for their observation is shown in cross section in Figure 86 and consists of the sections I to IV. These may be slid over one another along the planes *a-a′*, *b-b′*, and *c-c′*. Through the cell runs a U-shaped channel *d-d′* of rectangular cross section. To form a boundary, the channel is filled with the buffer solution of protein to a level slightly above the plane *b-b′*. Section III is

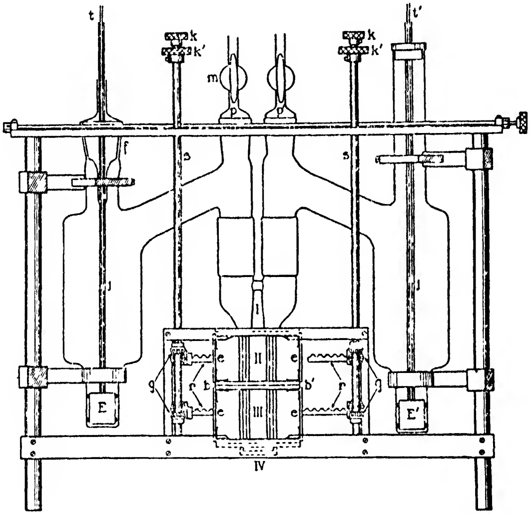


FIG. 87. Support and electrode vessels for Tiselius' cell.

then pushed to one side, the excess solution in section II is removed, and this section is rinsed with buffer. The remainder of the cell and the attached electrode vessels are then filled with the buffer. The support for the cell and the electrode vessels, together with the mechanism for moving the sections of the cell in relation to each other, as used at the Rockefeller Institute, are shown in Figure 87. Turning the knurled knobs *k-k′* operates bevel gears which in turn cause a horizontal motion of the racks *r*. Each rack presses against a metal insert, which communicates pressure to the edges of the horizontal glass plates. The sections

II and III of the cell may therefore be shifted in either direction by manipulation of the appropriate knob. The electrodes *E* and *E′* serve to lead current in and out of the apparatus. The large chambers *j*, filled with buffer solution, keep the electrode products from entering the channel in which the moving boundaries are being observed.

One of the most interesting uses of the electrophoretic method is the determination of the mobilities of proteins as functions of the pH and the ionic strengths of the buffer solutions in which

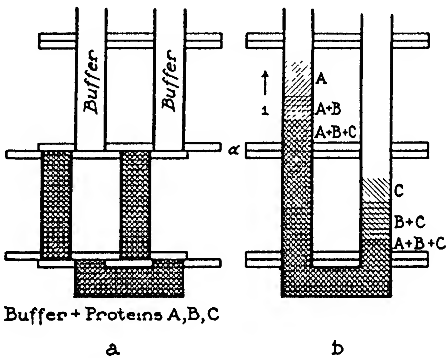


FIG. 88. Diagrammatic representation showing formation and motion of electrophoretic boundaries of a mixture of proteins.

they are dissolved. Some of our results will be given below. In addition to the measurement of mobilities the method furnishes information as to the purity of a protein in solution. It can also be used for the determination of the number (and, as we will see later, the concentrations) of the components of a mixture. This latter application is illustrated diagrammatically in Figure 88. Suppose a mixture of proteins *A*, *B*, and *C*, for which the mobility of *A* is greater than that of *B*, and that in turn is greater than the mobility of *C*, is placed in the cell as shown in Figure 88, a. Movement of the bottom center section of the cell to the right will bring the protein and buffer solutions into contact in

the plane α. On passage of a current three boundaries will appear in the upper cathode and in the lower anode sections, as is shown in Figure 88, b. It is evident from the figure that, in the ideal case assumed, there is a separation of pure component A in the region between the two leading boundaries on the cathode side and of pure component C between the two slowest boundaries on the anode side.

However, since protein solutions are, in most cases, colorless, the boundaries that form are not visible without special optical arrangements. The most convenient and successful method for studying the boundaries is based upon Toepler's schlieren

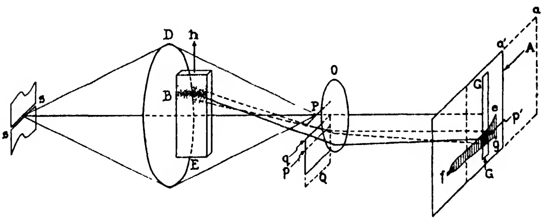


FIG. 89. Diagram of Toepler's schlieren system and of the schlieren-scanning method.

method. The principle of the schlieren technique is shown in Figure 89 and may be described as follows. An image at P of the illuminated slit S-S is formed by the schlieren lens D. The camera lens O is focussed on the electrophoresis cell E and forms an image on the screen at G-G. Now if the fluid in the cell E is homogeneous this image will be uniformly illuminated. On the other hand, if there is a boundary, B, between, for instance, a protein-bearing solution and a buffer, there will be a region in which the refractive index varies with the height, h, in the cell, and light which would normally pass to P is deflected downward, since the solution with the greater refractive index is on the bottom. If the opaque schlieren diaphragm Q is raised to a point p where it intercepts the most deflected light, a dark band will appear on the screen, corresponding to the region of the most

rapid change of refractive index in the boundary B. Such a band is shown in the image G-G at p'. If there is more than one boundary there will be more than one band. Figure 90, from a paper by Landsteiner, Longsworth, and Van der Scheer [20] shows the schlieren bands observed during the electrophoresis of a mixture of the albumins from duck and guinea hen eggs. At the commencement of the process a single band was seen. However, as current was passed this moved up the channel of the cell and split into two boundaries which progressively separated from each other, indicating two components with different electrophoretic mobilities.

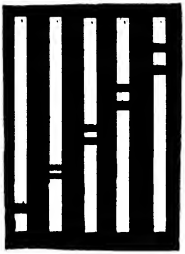


FIG. 90. Schlieren bands obtained during the electrophoresis of a mixture of albumins from duck and guinea hen eggs.

On the other hand, a boundary between two solutions does not consist of a single geometric plane, but of a region in which the composition varies from that of one solution to that of the other. Thus the refractive index changes continuously with the height in the cell from point to point in such a region. As we shall see, more important conclusions may be drawn from a complete knowledge of the distribution of the gradients of refractive index through the region including the boundary or a series of boundaries than can be obtained from the mere presence or absence of schlieren bands. The available methods for obtaining the complete information are the scale method of Lamm,[21] the diagonal schlieren method of Philpot and Svensson,[22] and the schlieren-scanning method [23] developed in the Rockefeller Institute laboratory. The results here outlined have all been obtained with the latter method, and it is the only one I shall describe. The principle is also indicated in Figure 89, in which the variation of the gradient of the refractive index with the height for a single boundary is represented

20. Landsteiner, Longsworth, and Van der Scheer, 1938.
21. Lamm, 1937.
22. Philpot, 1938; Svensson, 1939.
23. Longsworth, 1939; Longsworth, Shedlovsky, and MacInnes, 1939.

by the density of the shading. The pencil of light passing through the layer having the maximum value of the gradient will be most bent from the normal path and will be the first to be intercepted when the schlieren diaphragm *Q* is raised. On lifting the diaphragm still further, less refracted pencils of light will also be intercepted and the schlieren band in the image *G-G* will broaden. This can obviously be carried out until the whole field has been covered. In the schlieren-scanning method this process is made continuous and is recorded photographically. The image of the cell *G-G* is masked by a narrow vertical slit, and a photographic plate *A* is moved in the direction of the arrow at a constant rate across the slit. Actuated by the same mechanism, the schlieren diaphragm is given a steady movement upward. The resulting (positive) photographic record for a typical single boundary is indicated by the area *e-f-g* of the figure.

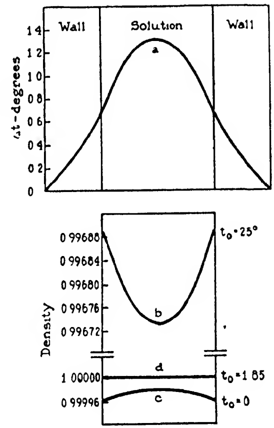


FIG. 91. Temperature and density relations across a tube through which current is passing.

As already mentioned the early study of electrophoresis by the moving-boundary method was made difficult because the boundaries were disturbed by thermal convection. The recent advances in the use of the method are largely due to a simple but brilliant idea of Tiselius. If an electric current is passed through a tube containing a conducting solution the distribution of temperature inside the tube and in the wall of the tube will be somewhat as is shown by curve *a* in Figure 91, i. e., there will be a maximum of temperature at the center of the tube. At ordinary temperatures and with aqueous solutions this means that there will be a minimum of density of the solution in the middle of the tube as is indicated, for the special case of a .1 normal acetate buffer, in

curve *b* of the same figure. As a result the heavier solution around the edges of the tube will tend to fall, and that in the middle to rise, and a convection current will result. However, although it is not possible to eliminate the temperature variation if the current is passed, one can choose a temperature at which the *density* variation is a minimum by working at a temperature near that of the maximum density of the buffer, which, for the solution under consideration, is 2.85°. If the thermostat temperature is such that the mean temperature in the tube is 2.85° the variation of density across the tube is shown by the straight line *d* of the figure. Theoretically, and also experimentally, this condition results in the substantial elimination of the effects due to convection. As shown by curve *c* of the figure, the variation of the density across the tube can be reversed, in comparison with that at room temperature, if the temperature of the thermostat is zero degrees.

The equipment for electrophoretic measurements at the Rockefeller Institute is shown in Figure 92, and consists of two separate instruments. In the rear of the room are the light sources. Next come the two thermostats, which operate at zero degrees centigrade. The two long tubes are the cameras, which are fitted with the scanning device already described. Each instrument is provided with a panel with the control switches and the means for measuring current and time. To reduce the disturbing effects of vibration, the optical bench, made from heavy steel girders, is mounted on concrete blocks resting on the floor of the basement laboratory.

The Electrophoretic Study of Ovalbumin and Egg White

Using the apparatus as described, it is possible to obtain "electrophoretic patterns" of proteins and of other substances. Figure 93 shows, for the rising and descending boundaries, the electrophoretic patterns of ovalbumin, the principal albumin of egg white. The material had been purified by successive crystallizations. It will be observed that each pattern shows two peaks, indicating two constituents with different electrophoretic mobilities.

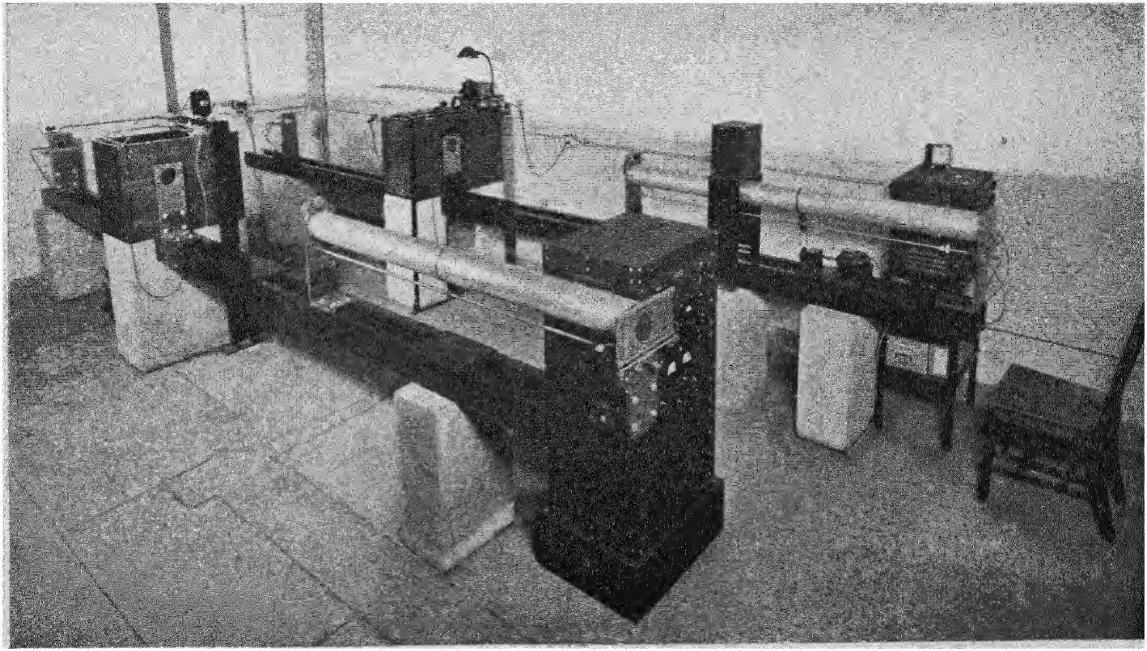


FIG. 92. Electrophoresis equipment in the Rockefeller Institute Laboratories as described on page 222.

This was somewhat disturbing because crystalline ovalbumin has been considered by many workers to be a single pure substance. However, material from a number of sources and prepared in several ways always showed the presence of this extra component. It is an interesting and important fact that, if an appropriate constant is known, the amounts of each component present may be found by determining the areas under each peak. The percentage of the second component may thus be estimated to be about 25 per cent.

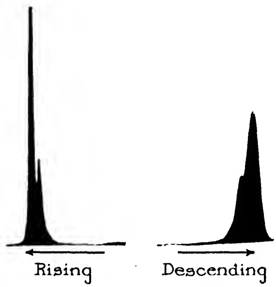


FIG. 93. Electrophoretic patterns for rising and descending boundaries of ovalbumin.

By an adaptation of the electrophoretic process the various components of a mixture may be separated. However, up to the present it has been found more convenient to use the electrophoretic patterns to study and control the usual methods of separation. An example of this is shown in Figure 94 which shows electrophoretic patterns obtained during the various stages of the purification of ovalbumin. Of these patterns, *a* of Figure 94 represents the result obtained, for the rising and descending boundaries, with egg white at pH 4. It will be seen that there are five peaks corresponding, as we have learned by experiments with the pure substances, to ovalbumin, A (which at this pH shows only one peak); conalbumin, C; ovomucoid, O; and the globulins, G_1 and G_2. (A peak for an additional globulin, G_3, is hidden under that of the ovalbumin.) The ovomucoid peak is complicated by the presence of δ and ϵ boundaries, for which no corresponding components exist. If now the conventional separation of the egg white is made into albumin and globulin, by precipitation of the latter with ammonium sulphate, the two fractions give the patterns shown in *b* and *c* of the figure. It will be seen that the albumin fraction differs but little from that of the egg white from which it has been separated, although one of the globulin peaks is

missing. Also that the globulin fraction, *c*, contains all the components of egg white, though in very different proportions. It is evident that the ammonium sulphate treatment does not produce a clean separation of any of the components. However, if the albumin fraction is recrystallized an effective purification of the ovalbumin begins. In pattern *d* of the figure obtained after one crystallization, it will be observed that there is no evidence of globulin, and the proportion of conalbumin is much diminished. With two more crystallizations, giving the patterns *e* and *f*, the conalbumin has, within the accuracy of the method, disappeared. It will be observed that the "anomalous" boundaries, δ and ε, persist in all the patterns. From this and other examples it appears likely that electrophoretic control, such as this example has just outlined, will be of service in the future, in a great many fields of endeavor.

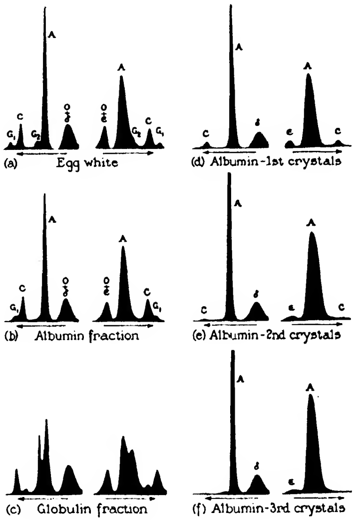


FIG. 94. Electrophoretic patterns obtained during purification of ovalbumin as described on page 224.

It seems appropriate at this point to consider the nature of the conductance by a solution of a protein. The present ideas are all grouped about the notion that protein molecules are ions, but of a very special kind. Their mean net charges are not constant but depend upon the acidity or alkalinity, i. e., the pH values, of the solutions in which they are placed. In acid solutions they take on positive charges, increasing to a maximum as the acidity is increased, that is to say, the pH is lowered. In alkaline solutions they assume net negative charges, and move to the anode in an electric field. Each protein has an *isoelectric point* which is the

pH value at which the protein in solution moves neither to the anode nor to the cathode when a potential is applied. The evidence is all in favor of the assumption that the protein is not uncharged at this point, but that the positive and negative charges balance each other. The protein in this condition is called a "*zwitterion*," or, in English, a dipolar or multipolar ion. A protein ion thus differs from, say, a sodium ion, in that its mobility, or rate of motion in a unit electric field, is a function not only of

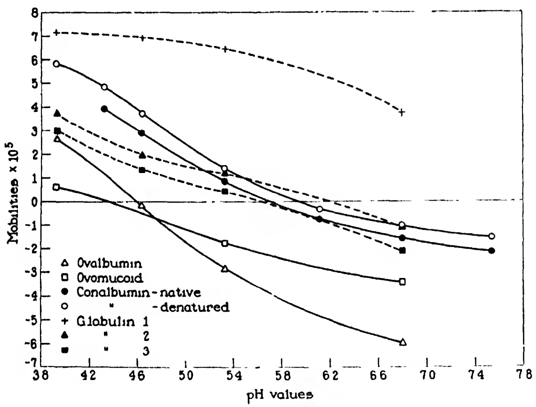


FIG. 95. The variation of the mobilities of the constituents of egg white with the pH value.

the ionic strength of the solution, but also, very markedly, of the pH of the solution. It is from these amphoteric properties of proteins that much of our knowledge at least of the external surfaces of the substances is derived and it is a method of study which will certainly be much used in the future.

With the aid of Dr. R. Keith Cannan, of New York University, Dr. Longsworth and I have made an investigation of the electrophoretic properties of egg white. The results are shown in Figure 95 in which the mobilities of the constituents (at an ionic strength of .1) are plotted as functions of the pH value. It will be seen that the mobilities of the constituents decrease with increasing

pH, passing through zero (at their isoelectric points) and then change sign. One globulin, G_1, however, is so basic in nature that it has no isoelectric point in the pH region studied. The data given in Figure 95 have been obtained from the purified constituents, with the exception of the globulins, which have been separated as a group but not as individuals. The mobilities at each pH value have also been obtained from the patterns, such as is shown in Figure 94, *a*, of egg white itself, and agree, in general, when the viscosity of the solutions is taken into account.

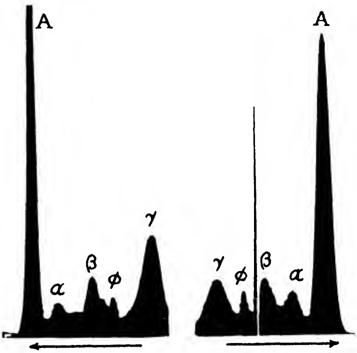


FIG. 96. Electrophoretic patterns of human blood plasma.

These studies have revealed several types of protein complexity. One, the presence of a "satellite" in ovalbumin, has already been mentioned. Another type of complexity is shown by conalbumin. The purified material has two components, which have tentatively been called "native" and "denatured." As shown in Figure 95, these have parallel mobility curves. However, there is a gradual shift of the proportions of the two components, the native being in preponderance at high pH values, and disappearing slowly as the pH is reduced. So far, only the native form has been found in egg white.

At high pH values no evidence of the globulins is obtained from the electrophoretic patterns of egg white. The patterns indicate, however, that under these conditions there is loose combination between the ovomucoid and the globulins. From this and other observations it appears that the electrophoretic method may yield important information concerning intermolecular combinations of the proteins.

The Electrophoretic Study of Blood

Of the protein-bearing fluids the most interesting and important would appear to be human blood. Stenhagen [24] has obtained the mobilities of the constituents of human blood plasma as functions of the pH by the schlieren-band method. However, by the schlieren-scanning method it has been possible to obtain not only the mobilities but also the concentrations of the constituents in the sera and plasmas. A brief summary may be given of the work done under my direction in the Rockefeller Institute laboratories by Drs. Longsworth, Shedlovsky, and Scudder.

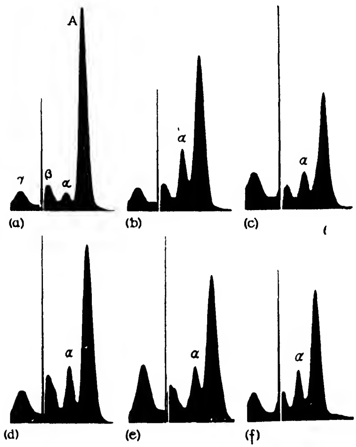


FIG. 97. Electrophoretic patterns of normal and pathological human blood sera. (a) Normal; (b) Pneumonia; (c) Peritonitis; (d) Peritonsillar abcess; (e) Rheumatic fever; (f) Lymphatic leukemia.

Figure 96 shows the electrophoretic patterns for the rising and falling boundaries of normal human plasma at pH 7.8. The peaks for albumin A and the three globulins, α, β and γ, as well as that for fibrinogen, ϕ, are clearly evident. The two patterns are roughly mirror images of each other except that the peak for the γ globulin includes the δ boundary in the pattern for the rising boundaries, and the much smaller ϵ effect for the descending boundaries. Also, there is, in the latter case, the interesting "β globulin effect," indicated by the sharp spike. The explanation appears to be that when separated, by electrophoresis, from the albumin and α globulin a change takes place in the β globulin, producing convection cur-

24. Stenhagen, 1938.

rents and, in consequence, regions of total reflection. In some cases the β globulin precipitates, yielding a turbidity which travels with the boundary. It is our experience that the patterns obtained from the blood of healthy individuals are strikingly reproducible.

In the course of the work at the Rockefeller Institute the electrophoretic patterns of the plasmas and sera of patients suffering with a wide variety of diseases were studied. Accounts of a part of this work have already been published.[25] Some of these electrophoretic patterns differ so much from that of normal blood that any interpretation was at first quite difficult. However, order has begun to emerge, and tentative conclusions as to the meanings of the patterns may be made. In Figure 97 the patterns obtained from the sera of bloods of patients suffering from various pathological conditions are compared with that from normal serum. One thing common to all the pathological conditions was fever. It will be observed that in each case there is a marked increase of α globulin. This increase of α globulin with fever has been found without exception.

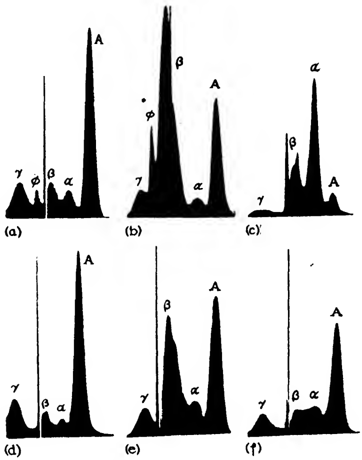


FIG. 98. Electrophoretic patterns of normal and pathological human blood plasmas and sera. (a) Normal plasma; (b) Multiple myeloma plasma; (c) Nephrosis serum; (d) Aplastic anemia serum; (e) Obstructive jaundice plasma; (f) Obstructive jaundice plasma (ether extracted).

Figure 98 contains the electrophoretic patterns of other samples of pathological sera and plasmas, including that of normal plasma, *a*, for comparison. Of these patterns, *b* is that of the plasma of a multiple-myeloma patient. The most marked

25. Longsworth, Shedlovsky, and MacInnes, 1939; Longsworth and MacInnes, 1940.

difference between this pattern and that of normal plasma is the great increase of the peak due to β globulin, or more exactly, of material moving with the mobility ascribed to β globulin, since a peak may arise from two or more materials moving with essentially the same velocity. It is our observation that lipoid material may be carried with the β globulin, probably in loose chemical combination. This is indicated by the patterns *e* and *f* of the figure from a case of obstructive jaundice. The first of these is directly from the plasma, whereas the second was obtained from the plasma after it had been extracted with ether to remove the fatty material. It will be observed that the peak due to the β globulin has been much reduced by this procedure.

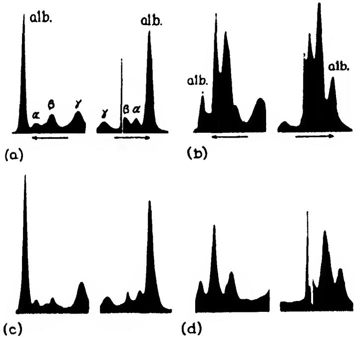


FIG. 99. Electrophoretic patterns of nephrotic sera and urine. (a) Normal serum; (b) Nephrotic serum; (c) Nephrotic urine; (d) Nephrotic serum (ether extracted).

As already mentioned, the concentrations of the various constituents appearing on an electrophoretic pattern may be obtained by measuring the areas under the various peaks. It will be observed that the pattern for the serum of aplastic anemia, Figure 98, *d*, is very close to that of the normal. However, by a quantitative study of this and other patterns of sera of aplastic anemia it was found that the ratio of albumin to globulin for such cases is decidedly below the normal.

Although I have no examples to show from our own work, it has been found that the antibodies, associated with immunity to diseases, travel during electrophoresis with or near the γ globulin.

Possibly the most striking result obtained in the electrophoretic study of blood and body fluids is that connected with nephrotic serum and urine. The patterns are shown in Figure 99. It will be

seen that the nephrotic serum *b* differs widely from the normal. Once more it is evident that the large increase of β globulin is due to lipoids, since the β globulin peak can be reduced by ether extraction, the result of which is given in pattern *d*. However, the patterns of nephrotic *urine*, *c*, resemble closely that of normal *serum*, with the exception that there is no β globulin disturbance.

CONCLUSIONS

As stated in the introduction, one of my purposes has been to show the utility of one type of experimental technique, that of moving boundaries, in different fields of scientific research. It has led us through the rather abstract notions concerning the behavior of electrolytic ions in solution, to the concentrations of protein constituents of blood in health and disease. However, these fields are not as far apart as might, at first thought, be supposed. The present attack, by physical chemists, on the constitution of proteins is largely toward an explanation of the variation of protein mobilities with pH, such as is shown in Figure 95, and the corresponding variations with the ionic strength. Work along these lines is going on in a number of laboratories, and must shortly yield important results. The guiding idea in all these investigations is that the interionic attraction theory of Debye and Hückel, briefly outlined here, also holds, with necessary modifications, for proteins in solution, since, acording to our present ideas, they are present as large and complicated ions. Further knowledge of the physical chemistry of proteins will certainly be followed by a better understanding of the physiological functions of the proteins in the blood and the other body fluids.

IX

THE ULTRACENTRIFUGE

By J. W. Beams

University of Virginia

Since the dawn of civilization man probably has been more or less conscious of the fact that larger particles settle through a liquid faster than smaller ones. In fact, he undoubtedly used this phenomenon in his crude mining and manufacturing operations. However, it remained for Galileo about 1590 to state clearly the fundamental laws underlying the phenomenon and to establish them by logical experimental procedure. From the equation established by Stokes [1] in 1847 it is clear why the process of settling in liquids is so effective for the separation of particles of different sizes, since, if the particle diameter is doubled, its speed of fall in the liquid is increased fourfold. If this equation of Stokes described the only phenomenon taking place in the process of sedimentation, one should expect, after sufficient time had elapsed, to find all uncharged particles or molecules completely settled out of suspension or solution. However, when the particles are very small, this is not the case because of another phenomenon, known as diffusion, which opposes complete sedimentation. This phenomenon of diffusion arises from the fact that all particles have thermal agitation or Brownian motion and are moving in random directions in the liquid.[2] Diffusion always

1. Stokes showed the theoretical relation between the rate of fall, v, under the force of gravity of a spherical particle of density ρ_p and its radius r in a liquid having a density ρ_d and coefficient of viscosity η to be

$$v = \frac{2}{9\eta}(\rho_p - \rho_d)gr^2.$$

2. Each particle has an amount of energy equal to $3/2\ RT/N$, where T is the absolute temperature, R the gas constant per mole, and N is the Avogadro number.

operates in such a manner that there is a net transport of particles from a region where the particles are closer together to where they are further apart. Consequently, in the process of sedimentation the settling out of small particles, which concentrates the particles at the bottom, will proceed only to the point where the sedimentation is balanced by diffusion. When the size of the particles approaches that of molecular dimensions, the diffusion is so great in comparison to the velocity of sedimentation that practically no particles will settle out under the force of gravity, unless, of course, the solution is completely saturated. The quantitative mathematical theory for the settling of particles in liquids under the force of gravity has been worked out by Mason and Weaver. The variation in the density of the air with height above the surface of the earth is an example of the process of sedimentation of a gas.

If, instead of allowing the small particles to settle out of a liquid or gas under their own weight in the gravitational field of the earth, the liquid or gas containing the particles is placed in a centrifuge and subjected to a centrifugal force many times that of gravity, the sedimentation becomes more complete. In addition to producing a greater amount of sedimentation, the larger centrifugal field makes possible the separation of particles with smaller size or mass differences. Because of these advantages of the centrifuge, it has been widely used, both in industry and in research. In this chapter an attempt will be made to describe some of the recent important developments in high-speed centrifuging, as well as some of their applications to research.

ULTRACENTRIFUGES

Recent progress in the technique of high-speed centrifuging has been made in two directions: first, the rotational speed of the centrifuges, and hence the centrifugal force, has been increased so that the rate of sedimentation is increased; and, second, the centrifuge has been made convection free so that no remixing of the material being separated can occur. Although it might seem

obvious that the latter development would always accompany the former if genuine progress is to be made, it was not until about 1924 that Svedberg first succeeded in obtaining convection-free sedimentation in centrifugal fields as high as 5,000 times gravity. In this pioneering work of Svedberg and his students it was shown that the centrifuge could be used effectively for the determination of particle or molecular size and shape. The results obtained with this early machine, which they named the ultracentrifuge, especially in the determination of the molecular weights and sizes of proteins, were of such great interest and fundamental importance that Professor Svedberg set about to improve the centrifuging technique still further. The results of this systematic investigation are the two modern types of "Svedberg ultracentrifuges," adapted especially to Professor Svedberg's epoch-making studies of particle or molecular weights and sizes.

The first type of Svedberg ultracentrifuge gives centrifugal fields from approximately 500 to 15,000 gravity. It is provided with ball bearings and is driven by an ordinary electrical motor. The rotor spins in hydrogen at atmospheric pressure and is surrounded by a water-cooled casing. The material to be centrifuged is contained in a sector-shaped cell with quartz or glass windows so that the concentration of the material at various radial distances in the cell can be determined by optical means while the rotor is spinning. This low-speed machine is usually used for the determination of particle or molecular weights by the so-called equilibrium method, which will be described later. The second type, or "oil turbine" Svedberg ultracentrifuge, is his famous high-speed machine which is used to give centrifugal forces in the range from 15,000 to 750,000 g. The rotor of this machine is made of a nickel steel alloy and shaped for maximum strength. It carries a sector-shaped cell with crystal quartz windows for observing optically the concentration of the material being centrifuged. The rotor is supported in horizontal bearings and is spun by twin oil turbines, one on each end of the rigid shaft. The oil

under pressure which drives the turbines is supplied by a special oil compressor and is filtered and thermostated at a suitable temperature before striking the turbines. The rotor is surrounded by a thermostated, heavy steel case. Hydrogen is continuously admitted at the periphery and pumped off at the center at such a rate as to maintain a pressure of about 20 mm. surrounding the rotor. The purpose of this hydrogen is to conduct the heat generated in the bearings by the oil impinging on the turbines and by the gas friction on the rotor to the casing, thus preventing the temperature of the cell from changing or becoming non-uniform. The rotor containing the cell and its contents is carefully balanced, both statically and dynamically, before running the machine. This machine is used principally for the determination of molecular weights and sizes by the rate of sedimentation method, which will be described later. Although the leading role which the Svedberg ultracentrifuges have played in the fundamental experiments of Professor Svedberg and his collaborators assuredly entitles them to a very detailed description in any analysis of high-speed centrifuging, my personal knowledge of them is secondhand, so I shall confine my discussion to the type of centrifuges and their uses which are more familiar to me.

Air-Driven Centrifuges

In general there are two different types of air-driven ultracentrifuges in use at the present time. The first type spins in air or other gases at approximately atmospheric pressure, while the second, or vacuum type, spins in an air-tight chamber which may or may not be highly evacuated. For most problems the second, or vacuum type, ultracentrifuge is definitely much superior, yet the first type still is often used because of its extreme simplicity of construction.

It has long been known that a ball may be supported and spun on a jet of air. However, it was not until 1925 that Henriot and Huguenard succeeded in constructing small rotors that could be

supported and spun to very high rotational speeds by properly directed jets of air. The original design of Henriot and Huguenard has been modified, improved, and stabilized by a number of different workers until at the present time the rotors are very stable. Figure 100 shows a diagram and Figure 101 a photograph of one of these simple air-driven, air-supported centrifuges which

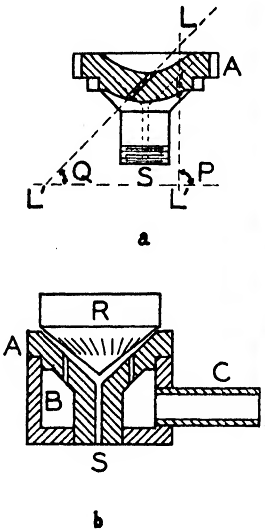


FIG. 100. Diagram illustrating air-driven, air-supported centrifuge as described on page 236. (a) Section through the strator cone (noncentral); (b) Section through complete machine.

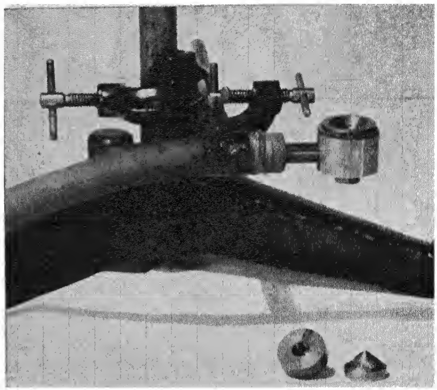


FIG. 101. Photograph of air-driven centrifuge shown in Fig. 100, mounted in clamp stand.

we designed and which has operated very satisfactorily in our laboratory for about ten years. The machine consists of three parts: a so-called stator cone A, an air box B, and a cone-shaped rotor R. Figure 100, b, shows the first two screwed together and the rotor in place. Parts A and B may be made of any machinable

material such as brass or duralumin, but the rotor R should be constructed of alloy steel or duralumin ST14 and shaped to give it a maximum bursting strength.

The construction of this machine is so simple that often we have given it as a problem in machine-shop practice to first-term

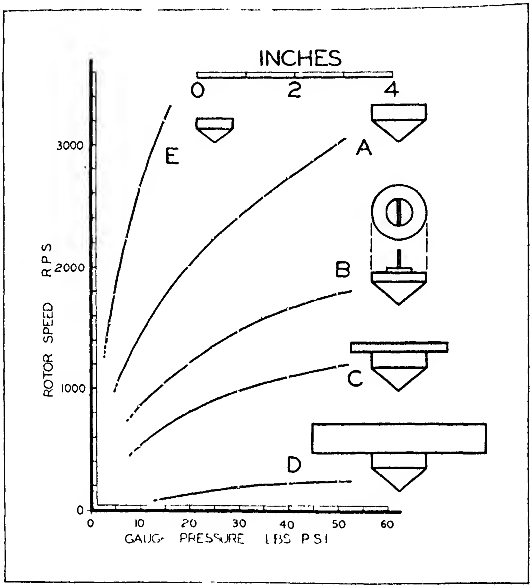


FIG. 102. Curves showing the relation of air pressure to rotor speed for rotors spinning in air at atmospheric pressure. The size of rotor for each curve is shown at right of the curve.

graduate students. To operate the machine, compressed air, that is forced into the air box B, emerges through the tubes LL[1] in jets which impinge upon the flutings of the rotor R. These air jets lift the rotor and start it spinning. The cone-shaped surfaces of A and R are so constructed that the Bernouilli forces prevent the rotor from flying out of the stator and compel it to spin on a thin cushion of air between them. The channel S allows air to

flow into the stator from the atmosphere and stabilize the supporting air cushion. This stable air cushion permits the rotor to seek its own axis of rotation within limits and therefore avoids the necessity of dynamically balancing the rotor with high precision. Also the rotor may be loaded or unloaded while at full speed or be made to carry a mirror or other superstructure. As a matter of fact, this type of machine has been used in our own laboratory principally for carrying a rotating mirror to photograph and observe phenomena that occur in very short intervals of time. It is not difficult to resolve times as short as a hundred millionth of a second with such a rotating mirror. Figure 102 shows the relation between rotor speed and the driving air pressure for a few typical rotors. It will be observed that as the diameter of the rotor is increased, the rotor speed is greatly decreased. This is due, of course, to the increased air friction on the rotors at high peripheral speeds. However, for small diameter rotors it will be observed that the rotational speed can be made very great. The highest rotor speed so far obtained in our laboratory is almost a million and one-half revolutions per minute with a 9 mm. rotor driven and surrounded by hydrogen. In this case the centrifugal force was about eight million times that of gravity.

In centrifuging experiments where precise temperature control is not essential and where a large centrifugal field is necessary only over a small radial distance, this simple machine has been used very effectively. For example, H. W. Beams, R. L. King, Harvey, Guyer, McIntosh, Selbie, Gatenby, and others have used this type of machine in very interesting studies of the relative displacement of the different components of plant and animal cells produced by centrifuging. Harvey and Pickels have devised optical methods of viewing material with a microscope while it is being centrifuged. In both methods the material is illuminated by a very narrow filament of light and the optical parts are so arranged that the plane of the image of the material. observed moves or rotates only a very small amount during the

time of viewing. It might be noted that the focal length of the microscope objective can be decreased when the rotor diameter is decreased. Hence for the highest resolving power the rotor should be small. Although these methods of Harvey and Pickels have not been used very extensively, it is probable that the adaptation of very small rotors would render them more useful.

This type of simple centrifuge has also been used by a number of different workers for the sedimentation of small particles and molecules out of suspension or solution. However, in these experiments great care must be taken to maintain temperature equilibrium throughout the centrifuge "bowl." The necessity for this can be seen when one considers that in a vertical vessel of water an increasing temperature from top to bottom of a degree or so will cause convection currents because of the smaller density at the bottom. One of the principal factors in producing this mixing or convection is the product of the difference in densities and the acceleration of gravity. Consequently in a centrifuge filled with liquid, where the field may be many hundred thousand times gravity, the temperature must be held extremely uniform to prevent convection. Of course there are other forces which oppose convection, such as the increase in density toward the periphery due to pressure, but these are small enough to require an extremely uniform temperature in an ultracentrifuge. In the centrifuge of Figures 100 and 101 it will be observed that temperature gradients exist in the rotor due to the cooling of the expanding air jets on its undersurface and the heating resulting from air friction on its periphery. Several devices have been used to make these rotors convection free. Dr. E. G. Pickels and I have used a thin rotor cell with a transparent window for observing optically the rate of sedimentation of hemoglobin. Uniform temperature of the cell was maintained by a second thicker cell directly below it, containing the same liquid, which broke up the temperature gradients by convection. We also obtained convection-free sedimentation by using disk-shaped baffles coaxial with the rotor, spaced one just above the other in the centrifuge

cell. McBain and his co-workers also have used various devices to prevent convection in these centrifuges and have reported the successful determination of sedimentation constants with their "transparent ultracentrifuge," as well as in ones in which the concentrations in the centrifuge cell are determined after the centrifuge is stopped. McIntosh, Selbie, and others have placed their material in small glass tubes in the rotor in a manner somewhat similar to the technique of Elford, with good results. However, for the measurement of rates of sedimentation, and thus particle or molecular size and weight, a little consideration will show that the precision is increased as the centrifuge rotor diameter is increased. Mason and Weaver, as well as Svedberg and his collaborators, have shown that the ability of a centrifuge to resolve two molecular species with almost the same sedimentation constants is proportional to w^2rh, where w is the angular velocity, r the radius, and h the height of the cell. In a general way h can be made larger as r is increased, so this resolving power is roughly determined by the square of the peripheral velocity of the rotor.

Now it will be observed from Figure 102 that a very large amount of energy is required to spin a rotor of more than about an inch in diameter to high peripheral speeds because of the increased air friction on rapidly moving surfaces. Consequently, in addition to troublesome thermal gradients set up in the rotor, the diameters of the rotors are too small to give high precision in this problem. High peripheral velocities are also essential for the concentration of isotopes, while in other experiments it is undesirable for the field to vary too rapidly. Because of these limitations we undertook the development of a high-speed centrifuge about seven years ago that would spin in vacuo.

The Vacuum-Type, Air-Driven Ultracentrifuges

The first vacuum-type, air-driven ultracentrifuge constructed in our laboratories consisted of a large rotor "centrifuge" situated inside a vacuum-tight chamber. This large rotor was suspended from and driven by an air turbine similar to that shown in Figure

100, located outside and vertically above the chamber. The turbine and centrifuge were connected by a piano wire which was coaxial with their vertical axis of rotation and passed through a vacuum-tight oil gland which sealed the vacuum chamber. Since the piano wire shaft was flexible the large rotor could seek its own axis of rotation and could be spun through any critical speeds without difficulty. Also the small diameter shaft permitted low friction losses in the oil gland or bearing even at high rotational speeds. The vapor pressure of the oil in the oil gland which limits the vacuum attainable was very low (10^{-4} to 10^{-6} mm. for vacuum pump oils) so that the air or gaseous friction on the large rotor was negligible. In actual practice it was found that the efficiency of the air turbine was not changed appreciably by attaching the large rotor and shaft to it provided the residual pressure in the vacuum chamber was below 5×10^{-4} mm. As it is possible to explode the small turbine by its own centrifugal force, the only limiting factor on the rotational speed of this machine is the strength of the large rotor. Since in a vacuum there is no heat generated by the rotor, the temperature of the rotor not only remained extremely uniform but could be accurately controlled by thermostating the vacuum chamber. Consequently, convection-free sedimentation was easily obtained. This vacuum-type machine has been improved and adapted to many different special problems by a number of workers and has been found to be a very efficient machine.

Figure 103 is a sectional diagram and Figure 104 a photograph of a vacuum-type centrifuge which we have used in an attempt to concentrate the isotopes by the method of evaporative centrifuging. The rotating parts consist of the centrifuge C inside the vacuum chamber, the flexible hollow shaft A (hypodermic needle gauge 15), and the air turbine T. The air turbine T is supported by air entering at I and forming a cushion between the Bakelite cup and the underside of T. The turbine is driven by air entering the air box S through D and impinging upon the turbine flutings. The vacuum-tight oil glands G_1 and G_2 are supported in Neoprene rings and have good thermal contact with their supports. They are arranged so that

oil may be circulated through them, but this is very seldom necessary. Vacuum-pump oil is forced into them under a pressure slightly above the air pressure in I. In experiments on the separation of chlorine isotopes, about 14 cc. of liquid carbon tetrachloride is injected into the hollow rotor C. It is then spun to 1,560 r.p.s. and evacuated through the hollow shaft A. The carbon tetrachloride evaporates at the periphery, diffuses in the

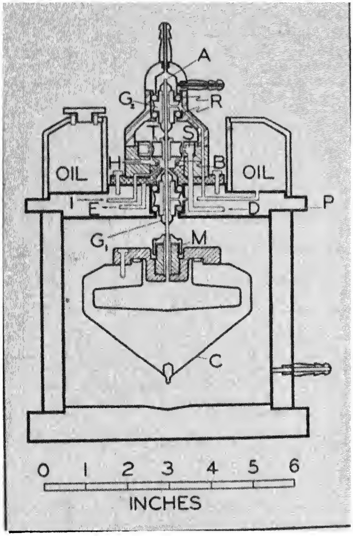


FIG. 103. Section of air-driven, vacuum-type centrifuge as described on page 241.

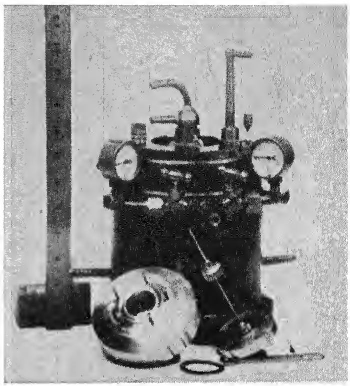


FIG. 104. Photograph of centrifuge shown in Fig. 103.

vapor state through the centrifugal field to the axis of rotation, out through the hollow shaft, and is collected in a succession of 2 cc. (liquid) fractions in dry-ice traps placed in the pumping line. It was found that the first fraction collected had the light isotope 35 concentrated, while the last samples had the heavy isotope 37 concentrated. When the rate of withdrawal was not large enough to destroy the equilibrium conditions in the rotor, the separation obtained was in accord with the equilibrium theory. It has been shown by Aston and Lindemann, Mulliken, Chapman, and others that for the case of a substance with two isotopes which behave as ideal gases, the

ratio K_0 of the concentrations of light to heavy isotopes at the axis divided by the same ratio K at the periphery

$$S = \frac{K_0}{K} = exp\,\frac{(M_2 - M_1)v^2}{2\,RT} \tag{1}$$

where S is the so-called separation factor, v is the peripheral velocity in cm./sec., $R = 8.3 \times 10^7$, and T is the absolute temperature. Mulliken has further shown that this same equation holds when the element is centrifuged in any of its compounds. It will be noted from the above theory that the separation factor S depends only upon the differences in masses of the isotopes and not their absolute value, so that the method should be practically as successful for heavy as for light elements. It will also be observed that the separation factor is greatly increased with increasing peripheral velocity v and decreasing temperature T. Table IV gives some values of S for peripheral velocities normally used in the apparatus in Figure 104 at room, dry-ice, and liquid-air temperatures:

TABLE IV

$M_2 - M_1$	v *cm./sec.*		S	
		300° *A*	200° *A*	90° *A*
1	4.5×10^4	1.04	1.06	1.13
2	4.5×10^4	1.08	1.12	1.30
3	4.5×10^4	1.12	1.19	1.47
4	4.5×10^4	1.17	1.27	1.68

Mulliken has also shown that in his evaporative centrifuge method used above, the change in atomic weight of the element ΔA is given by

$$\Delta A = \frac{(M_2 - M_1)^2 X_1 X_2 v^2}{2RT} \log_e C = B \log_e C \tag{2}$$

where X_1 and X_2 are the mole fractions of the light and heavy isotopes and C is the so-called cut or the ratio of the total amount of material initially to the amount of residue remaining in the centrifuge. Figure 105 shows a graph of ΔA versus the cut.

In order to take advantage of the increased separation factor at reduced temperature, the apparatus was modified to permit

the centrifuging of materials at other than room temperature. Essentially it is the same machine, but with the large rotor insulated thermally from the driving mechanism. The shaft is a long stainless steel tube, and the guides to keep it from developing standing waves at certain frequencies are mounted on nonconducting Bakelite. The large rotor is surrounded by a metal

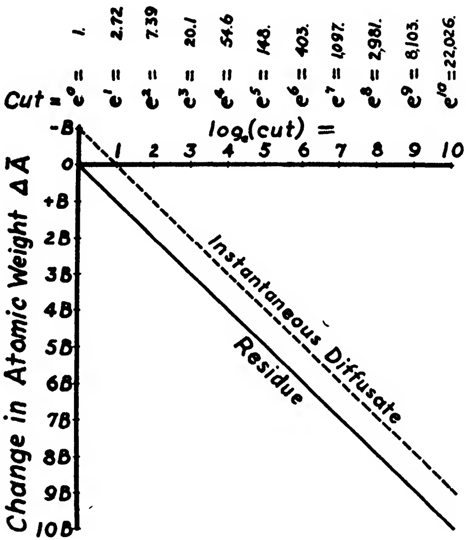


FIG. 105. Relation of atomic weight change to the natural logarithm of the cut for the evaporative centrifuge method of concentrating isotopes.

"thermos flask." With this apparatus the centrifuge has been spun at liquid air temperature, and the separation of the isotopes of chlorine in ethyl chloride at dry-ice temperatures has been shown to be in accord with the theory just given.

The practicability of any centrifuging process for the separation of isotopes depends not only on the separation factor S, but upon the amount of material that can be centrifuged per unit of time. The amount of material that can be centrifuged per unit of time depends upon the time required for rough equilib-

rium to be established between sedimentation and diffusion in the centrifuge (page 242). When the material is passed through the centrifuge too rapidly, the separation factor is decreased. This question has been treated theoretically by Humphreys and by Witmer and Wilson. In order to increase the amount of material centrifuged per unit of time, it was obviously necessary to increase the size of the centrifuge. A little consideration will show that it was more practical to increase the length than the diameter of the rotor. We therefore undertook the development of the technique for spinning long rotors (Fig. 106).

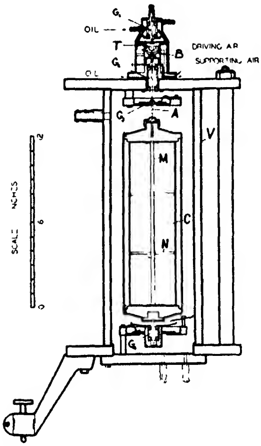


FIG. 106. Section of air-driven, vacuum-type tubular ultracentrifuge used in the evaporative centrifuge method of separating isotopes.

The only necessary change in the apparatus shown in Figure 103 is the flexible sliding bearing at G_3 and G_4. The long rotor is a chrome moly-steel tube approximately 1 foot long and 3 inches in internal diameter, with ½ inch wall thickness. The tube contains a large number of very small holes for drawing off the vapor uniformly along the axis of the tube.

Figure 107 shows a graph of the ratio of the separation expected from the theory to that observed versus the rate of removal of carbon tetrachloride per minute. The circles show the results ob-

tained with a hollow tube. It will be observed that the efficiency starts to fall off at rates of removal of about .2 cc. liquid CCl_4 per minute. This falling off was eventually traced to the remixing in the rotor at the high rates of withdrawal of the vapor. As the vapor moves from the periphery toward the center, its angular momentum is conserved and hence a sort of whirlwind is formed, spinning in the same direction as the rotor which gives

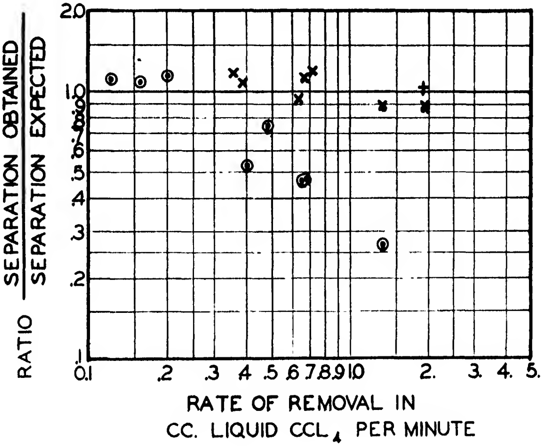


FIG. 107. Ratio of isotopic separation obtained to that expected on equilibrium theory versus rate of removal of CCl_4. The circles are for a hollow tube, and the crosses for the same tube with baffles.

rise to remixing. By putting spider baffles in the rotor to absorb the angular momentum of the gas and thus prevent stirring, the experimental values shown by the crosses in Figure 107 were obtained. The magnitude of the angular momentum given up by the gas in diffusing to the axis was demonstrated by the fact that after the baffles were installed the driving air to the turbine could be considerably reduced, yet the rotational speed of the centrifuge was maintained constant by the loss of the angular momentum of the CCl_4 vapor. This method of concentrating isotopes is very effective for changing the isotopic ratio by about

20 per cent or less in comparatively large quantities of material, but to get large changes in the isotopic ratio the process becomes very laborious.

In addition to the evaporative centrifuge method, we have also experimented with a continuous-flow method. Figure 108A shows a vertical section and Figure 108B a picture of the centrifuge used. The air turbine is placed below rather than above the

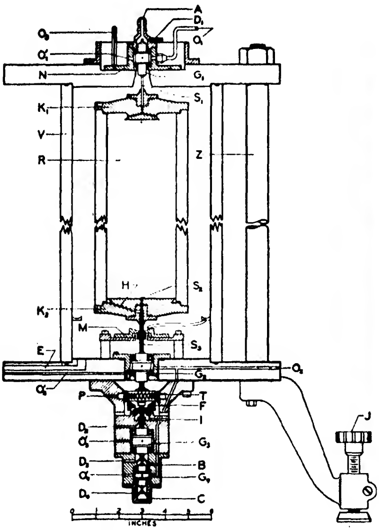


FIG. 108A. Section of continuous flow, air-driven, vacuum-type centrifuge used in the separation of gases and vapors as described on pages 247–248.

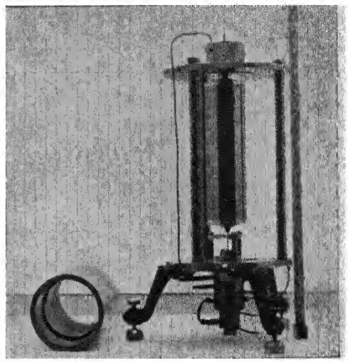


FIG. 108B. Photograph of centrifuge shown in Fig. 108A.

vacuum chamber; otherwise the driving and supporting mechanism is identical to that shown in Figure 103.

The tubular rotor R (4″ O.D., 15″ long, with ½″ wall thickness) is made of alloy steel and capped by duralumin end plates K_1 and K_2. The gas or vapor to be centrifuged flows at a continuous rate into A, through the stainless steel tube S into R, and as it flows down R separation takes

place. Upon reaching K_2 the heavy fraction flows out through the loosely fitting circular flange into the tubes H, then through the region between the coaxial tubular shafts S_2 and S_3 and is collected at B. The lighter fraction flows down S_2 and is collected at C. The ratio of the quantity of material collected at B and at C can be regulated by adjustable valves at B and C. In other words, the gases or vapors pass through the machine at a continuous rate and are collected in a heavy and a light fraction.

The machine has been used to separate nitrogen-oxygen, nitrogen-carbon dioxide mixtures, and the chlorine isotopes in ethyl chloride vapor. When baffles were placed in the centrifuge to prevent stirring, the ethyl chloride vapor could be passed through the machine at the rate of about 2 gms. per minute without appreciable reduction in the separation factor, which was in approximate accord with theory. It might be noted that this cylinder with its baffles weighed 36 lbs. and when spinning at 1,100 r.p.s. had stored in it something over a half million foot-pounds of energy.

The air-driven vacuum-type centrifuges, which I have been describing, were developed primarily for use in the concentration of isotopes. Since the isotopes differ usually by 1 or 2, and in rare cases, 10 atomic weight units, it is clear that they are, with the exception of compounds with almost equal molecular weight, among the most difficult of all substances to separate by centrifuging. Hence a centrifuge which gives even a small separation of the isotopes should be a very effective means of purifying many substances. This has been found to be especially true in the case of the very large molecular weight substances such as hormones, viruses, enzymes, etc., which are known to play an important role in our well-being. Many of these biological materials occur naturally in great dilution, and it is of first importance to concentrate and purify them. Often this is difficult, especially by chemical means, because small changes in pH value or temperature will deactivate them. Fortunately, however, their properties apparently are not affected by strong centrifugal fields, and the temperature of the types of centrifuges which I have

described can be controlled accurately. Hence these centrifuges are well suited for their purification.

Quantity-Type Ultracentrifuges

Although many different workers have used various modifications of the air-driven, vacuum-type ultracentrifuges in the purification of their materials, Figures 109 and 110 contain the general principles usually employed. The driving and supporting mechanism is essentially the same as that in Figure 103 except that additional provision is made for a set of reverse jets which may be used for decelerating the machine in a comparatively short time. The rotor C is a large metal block (usually duralumin ST14) machined to the approximate shape shown. The holes bored at an angle contain Lusteroid test tubes which, in turn, contain the material to be centrifuged. The vacuum-tight top is clamped to the rotor and the surrounding chamber evacuated to less than 5×10^{-4} mm. of mercury pressure. The rotor is then spun for the desired time and, upon stopping, the heavy materials are found to collect near the bottom of the tubes, whereas the lighter fractions are found near the top. Wyckoff has shown that, within limits, the smaller the angle which the axes of the tubes make with the vertical, the better the separation. An 8-inch rotor may hold 150 cc., while a 6-inch will hold 80 or 90 cc. In their first experiments Bauer and Pickels obtained a concentration of the yellow-fever virus of ten thousand times, whereas Wyckoff and Corey found that tobacco mosaic virus was crystallized on the bottom of the tube in one centrifuging.

This type of machine is well suited to the concentration of comparatively large molecular weight substances. However, since the time required to complete one centrifuging is usually several hours, the amount of material that can be centrifuged with one machine is very limited.

Recently the question has arisen whether or not the tubular vacuum-type centrifuge (Figures 107 and 108) used for gases and vapors might be used for the concentration of materials in

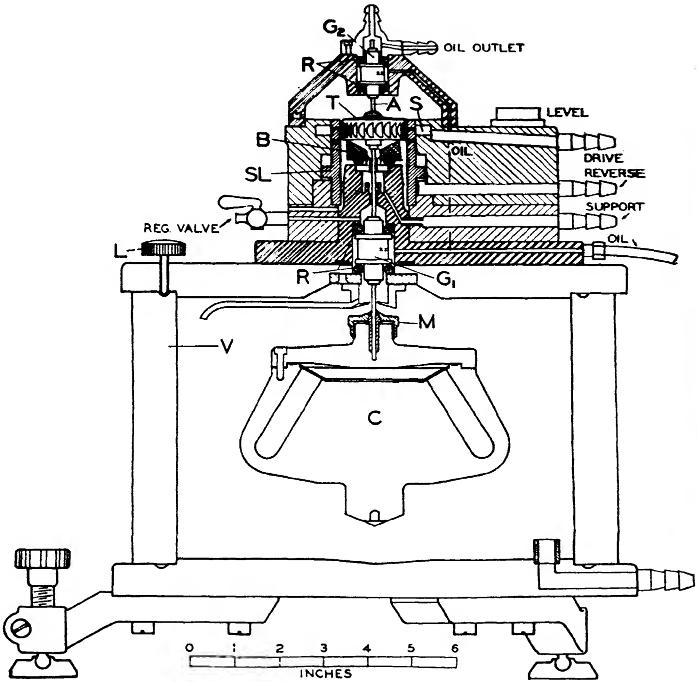


FIG. 109. Section of air-driven, vacuum-type centrifuge used for purification of large molecular weight substances as described on page 249.

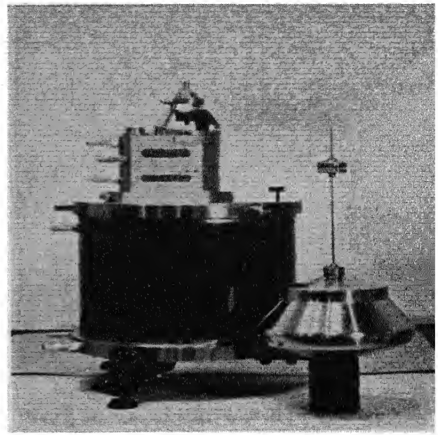


FIG. 110. Photograph of centrifuge shown in Fig. 109. The rotor-turbine and shaft are in the foreground.

solution. In collaboration with Drs. Chanutin and Masket, of the University of Virginia Medical School, we have been making some preliminary tests with a small tubular machine on a few biological substances such as egg albumen and hemoglobin. Several modifications in the machine of Figures 107 and 108 were, of course, necessary. A solid rod or core was placed inside the tubular rotor and coaxial with it, in order to force the materials being centrifuged into the strongest part of the centrifugal field, i. e., in the annular space between the core and inner wall of the tube where separation takes place. It was also necessary to make the rotor and core of stainless steel or else to plate them with gold to prevent corrosion. The inlet and exit chambers were made of Lucite. Guards were also placed between the inlet and exit chambers to prevent the material being centrifuged from coming in contact with the oil leaking from the vacuum-tight oil glands. With this arrangement the material flows into the centrifuge at the top at a continuous rate, and a heavy fraction and a light fraction are separately collected at the bottom in a manner similar to that of the cream separator. This machine is not yet fully developed, but our preliminary results show that the separation attained is in approximate accord with the predictions of the theory. This shows that no stirring or remixing is taking place. A little consideration will show that the maximum separation through this ultracentrifuge is proportional to the number of revolutions of the centrifuge per second squared, to the sum of the outer radius of the core and the inner radius of the tube squared, and to the length of the tubes. Since tubes several feet long can be spun up to their bursting point, this tubular ultracentrifuge for liquids holds considerable promise as a method of separating materials rapidly.

Analytical Ultracentrifuges

As previously mentioned, Svedberg and his collaborators have shown that the particle or molecular weights or sizes can be obtained from centrifuging data. There are two general cen-

trifuging methods, known as the equilibrium and the rate-of-sedimentation methods, used for molecular or particle weight and size determination at the present time. In the equilibrium method the centrifuging is continued a sufficient time for equilibrium between sedimentation and diffusion to be established at every point in the centrifuge.

When this equilibrium state is reached in a dilute solution contained in a sector-shaped cell,

$$M = \frac{2RT \log_e C_1/C_2}{(1 - \rho V) w^2 (r_1{}^2 - r_2{}^2)} \tag{3}$$

$$v = \frac{2RT \log_e C_1/C_2}{N(\rho_p - \rho_d) w^2 (r_1{}^2 - r_2{}^2)} \tag{4}$$

where M is the molecular weight, R the gas constant per mole, T the absolute temperature, C_1 and C_2 the concentrations at the points r_1 and r_2 distance from the axis of rotation, respectively, ρ the density of the solution, V the partial specific volume of the molecule, N the Avogadro number, ρ_p and ρ_d the densities of the particle and solvent respectively, and v the volume of the molecule.

The equilibrium method is the simplest from a theoretical standpoint and requires only the measurement of two concentrations at their respective distances from the axis of rotation in addition to the constants of the centrifuge and of the solution given in eq. (3) and (4) for a determination of the molecular weight or molecular volume. On the other hand, the time required for equilibrium to be established in a centrifuge is usually very long. According to Archibald, who has worked out the general theory for the sedimentation in centrifuges, as well as from experiment, this time required for equilibrium to be established may be several days in the case of comparatively large molecular weight compounds. Besides requiring a long centrifuging time, throughout which the temperature, rotational speed, etc., of the centrifuge must remain constant, the method cannot be used in the case of many of the most interesting (biological) substances be-

cause they will decompose before equilibrium can be established.

Fortunately, the rate-of-sedimentation method supplements the equilibrium method because it operates best where the equilibrium time is longest.

If the solution is ideal and dilute and if the particles or molecules are not electrically charged, and, also, if no particles are reflected from the ends of the sector-shaped sedimenting column, then the centrifugal force per mole can be set equal to the frictional force per mole—that is,

$$M(1 - V\rho)w^2r = f\,dr/dt \tag{5}$$

where the frictional constant $f = RT/D$ from Einstein's diffusion equation (D being the diffusion constant). Hence,

$$M = \frac{RT(dr/dt)}{D(1 - V\rho)w^2r} = \frac{RT}{D(1 - \rho V)}S \tag{6}$$

where S is the sedimentation constant which is the velocity of sedimentation in a unit field.

It will be observed that with this velocity-of-sedimentation method, it is necessary to measure both the sedimentation constant S and the diffusion constant D in order to determine the molecular weight. However, the measurements can be made in a short time if the centrifugal field and the molecular weight are large. The methods of measuring S (eq. 6) as well as C_1 and C_2 (eq. 3 and 4) usually are optical or analytical. The optical methods which have been developed by Svedberg and his students consist in determining the concentration of the material in a sector-shaped transparent centrifuge cell as a function of the distance from the axis of rotation while the centrifuge is spinning by the amount of light absorbed or by the refractive index. The index-of-refraction method has taken several forms such as the Lamm scale method, the schlieren method, etc., and is usually preferable except in cases where the material being centrifuged has convenient absorption bands.

A photograph of an ultracentrifuge rotor is shown in Figure 111, which contains a sector-shaped cell with crystal quartz windows which may be used for either equilibrium or velocity-of-sedimentation measurements. It is made elliptical in shape to increase its strength. The driving mechanism for this rotor is

the same as that shown in Figures 109 and 110. Figure 112 shows a series of photographs using the absorption method for determining the rate of sedimentation in diluted fresh human blood. It will be observed that, in addition to the downward motion of the boundary between the hemoglobin and the clear solvent, there is a slight blurring due to diffusion. With the aid of microphotometer measurements across these boundaries, it is not only possible to measure the sedimentation constant S but the diffusion constant D as well. However, D is usually measured in a separate experiment. If the solution contains more than one molecular species, then each will form a separate sedimentation boundary which moves out with its characteristic velocity. Also, if, instead of homogeneous molecular species, the solution contains a distribution of particle sizes or weights, this distribution can be determined by the "blurring" of the sedimenting boundary. For precise measurements the molecules or particles must be uncharged; otherwise the settling out of the heavier ions will set up electrical potentials which may oppose the sedimentation. This trouble is usually eliminated by centrifuging at the proper pH value or by adding a small amount of a low molecular weight electrolyte to the solution. Since the specific volumes used are based upon dry weight measurements, both the equilibrium and the velocity-of-sedimentation methods take no account of hydration effects but Svedberg and Kraemer and Lansing have shown that as long as the density of the hydration or absorption shell around the molecule does not differ appreciably from the density of the solution, the results are not affected. In fact, in cases of consid-

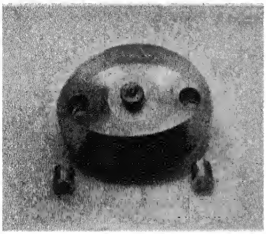


FIG. 111. Rotor used in equilibrium and rate of sedimentation methods for determining particle or molecular weight and size. Right foreground, cell with transparent crystal quartz windows. Left foreground, dummy cell necessary for balancing.

erable differences they show that the errors introduced are quite small provided the concentration of the protein and salt is small.

It might also be noted that the above methods can be used for determining particle or molecular sizes and shapes as well as volumes and masses.

In some cases it is not practical to measure the concentration as a function of the radius in the centrifuge by optical methods. Consequently it is necessary to make analyses of samples of the material taken from different distances from the axis of rotation of the centrifuge by the ordinary methods of analytical chemistry. These samples can be taken from the centrifuge while it is spinning or after it is stopped, provided no remixing takes place. Tiselius, Pedersen, Svedberg, and others have divided the sector-shaped centrifuge cell into approximately equal compartments by means of a partition which is perpendicular to the direction of sedimentation. The partition is usually made of thin Bakelite perforated with a large number of fine holes, which is covered with a piece of hard filter paper. This arrangement does not appreciably disturb the sedimentation but prevents remixing upon stopping.

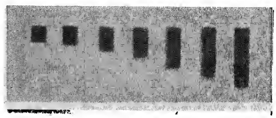


FIG. 112. Series of successive photographs illustrating rate of sedimentation of dilute human blood by the absorption method.

They have shown that using the velocity-of-sedimentation method, the sedimentation constant S is given by

$$S = -\frac{1}{2w^2 t} \log_e \left(1 - \frac{2\Delta}{q\, x\, c_0}\right) \qquad (7)$$

where Δ is the total change in amount of the material above or below the partition, x the distance from the axis of rotation, q the area of the partition, and c_0 the concentration of the substance at the beginning. However, this method is not very precise because for large molecular weight substances the time required to accelerate the centrifuge to full

speed and to stop it is an appreciable part of the running time t. Since the product $t\,w^2$ enters into eq. (7), a considerable error is introduced.

In order to increase the precision of this method, Mr. Fox and I have constructed the machine shown in Figures 113 and 114. The centrifuge is first accelerated to full speed and then the material to be centrifuged is introduced through the hollow

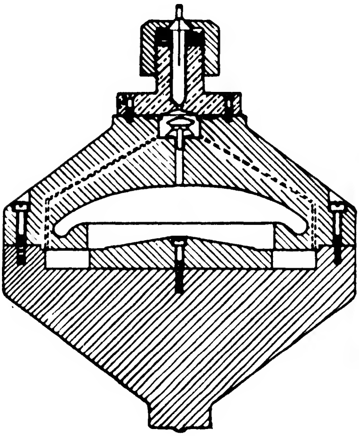


FIG. 113. Section of rotor which permits division of cell contents while machine is spinning, as described on page 256. The rotor is made in two parts and fastened together with screws.

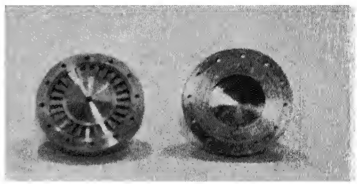


FIG. 114. Photograph of rotor illustrated in Fig. 113 with the two parts separated. Upper part, right; lower part, left.

shaft until the sector-shaped cells are all just filled. The centrifuging is then continued at a constant angular velocity w for the desired time t. A liquid heavier than the solution is then introduced through the hollow shaft. This liquid flows into the sector-shaped cells near their periphery and about half of the solution is forced into a chamber above the cells. After the rotor is stopped, the sedimentation constant is determined from analysis of the original solution and of the material collected in the upper chamber of the centrifuge. The rotor is spun by the same

driving mechanism as shown in Figure 104. The method has been used to measure S for egg albumen and for hemoglobin. It seems to be capable of high precision. Obviously the same apparatus can be used with the equilibrium method as well as the velocity-of-sedimentation method.

As mentioned before, it has been shown by Mason and Weaver for a uniform field and by Svedberg and his collaborators for the centrifugal field that the resolving power is equal to the product of the strength of the field and the height of the sedimenting column. That is, the ability of a centrifuge to separate the sedimenting boundaries of two homogeneous molecular species is proportional to the w^2rh, where h is the height of the cell, r is the radius, and w is the angular velocity. The quantity is roughly proportional to the strength of the rotor, which in turn is the limiting factor in all of the experiments described above. In the case of substances with very large molecular weights, such as tobacco mosaic virus, the whole resolving power of the centrifuge cannot be used because the material settles almost out of the cell before the machine can reach full speed. Recently we have been experimenting with a scheme by which the solvent is made to flow through the cell from the periphery toward the axis of rotation at about the same speed with which the material settles out. This makes it possible to keep the sedimenting boundary in the field of view for long periods of time, thus greatly increasing the resolving power of the machine. For example, tests with this machine show human hemoglobin to be very homogeneous. It might also be noted that in addition to increasing resolving power, this method increases the flexibility by making possible changes or variations in the solvent during the centrifuging. In addition to maintaining the sedimenting boundary stationary in the cell, by the above flow method it can be done electrically in some cases. By making two electrical connections to the two ends of the insulated sedimenting columns with insulated lead wires which pass out through the shafts and connect with mercury cups outside, it has been found possible to maintain

an electrical field in the cell in a radial direction. If the solution being centrifuged is properly buffered, and the substance is not at its isoelectric point, it is possible to hold the sedimenting boundary in the field of view by balancing the centrifugal force on the molecules by the force of the electrical field upon their charges. Some preliminary experiments show that this can easily be done. However, the flow method is far superior in both theory and practice.

The Electrically Driven, Magnetically Supported Ultracentrifuge

The air-driven vacuum-type machines described above are practically ideal centrifuges because they are convection free and will produce centrifugal fields as high as it is possible to construct rotors to withstand. Also rotors of almost any size can be spun successfully. However, they require a supply of compressed air, which is not always available. Also, if constant rotational speed is required, an automatic speed control which actuates the driving pressure is necessary. In order to obviate these inconveniences, the magnetically supported, electrically driven, vacuum-type ultracentrifuge has been developed. It employs the same mechanical principles as the air-driven vacuum-type machine except that the air support has been replaced by a magnetic support and the air turbine drive has been replaced by a small electrical motor. Figures 115 and 116 illustrate the machine which Dr. Skarstrom and I have found very successful.

The rotating parts consist of the armature of an electrical motor D, an iron rod R, the centrifuge C, and the flexible shaft A. The flexible shaft passes through the vacuum tight oil glands G_1, G_2, and G_3. The sliding guide at G_4 serves to prevent the rotor C from wobbling when it is first started. The rotating parts are almost but not quite supported by the upward pull of the electromagnet L on the iron rod R. The remaining small thrust is taken by the top bearing of G_3. Oil leaking from G_3 which is under pressure must pass out through this thrust bearing, so that the surfaces of the thrust bearing are separated by a film of oil. This bearing

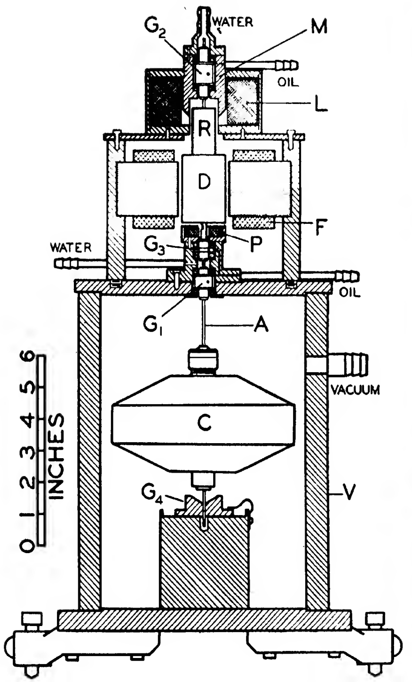


FIG. 115. Section of magnetically supported, electrically driven, vacuum-type centrifuge as described on page 258.

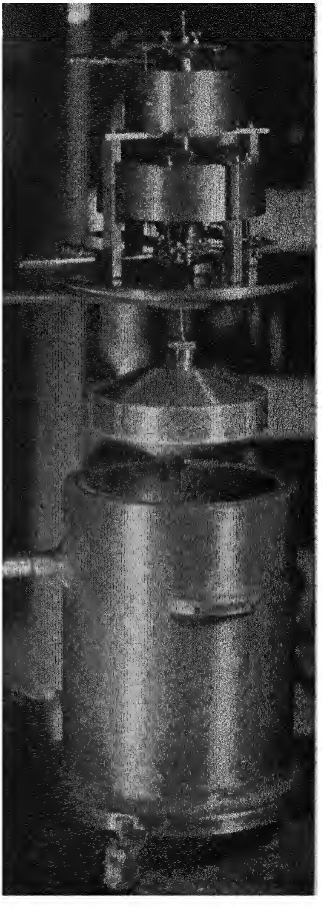


FIG. 116. Photograph of centrifuge shown in Fig. 115.

is able to support the whole weight of the rotating members for several hours without harmful wear in case the current in the solenoid L is interrupted. It, however, has considerably more friction than the magnetic support, which is almost frictionless because the symmetry of the magnetic field practically eliminates electromagnetic drag. The field coils of the motor are actuated by a high frequency alternating current in such a way that the field rotates with the same frequency as that of the alternating current. This induces currents in the solid steel armature D and causes the centrifuge to spin. The high frequency alternating current is generated by amplifying the power from a fixed or variable audiofrequency oscillator, feeding this into a network composed of two parallel condenser-inductance circuits, one tuned above the other below resonance, so that two-phase power is generated. This two-phase power is applied to the opposite pairs of the field coils F and produces the rotating field. The induced currents in D produce considerable heat during the starting period when the machine is accelerated rapidly, so that water is forced in at the top down through the hollow shaft to the region between G_1 and G_3, where it emerges from the shaft and flows out of the tube marked "water." The rotor C shown in the drawing and picture is 6¼″ in diameter and is made of solid duralumin ST14. It has a moment of inertia of 80,000 gm./cm.^{2} and weighs over 7 lbs. (the total weight of the rotating members is over 8 lbs.). When 1,000 watts are delivered to the motor at a frequency of 1,200 cycles, the centrifuge attains a speed of 1,060 r.p.s. in about 18 minutes. This acceleration period compares very favorably with the air turbine drive. To maintain the speed constant at any desired value (say, at 1,060 r.p.s. in the above case), a small magnetic pick-up at P delivers a signal voltage whose frequency critically controls the power input to the motor. This arrangement will maintain the rotational speed constant at 1,060 r.p.s. to $\pm$ .03 per cent for at least a day, i. e., to $\pm$ .3 r.p.s. The power input to the motor required to maintain this speed is about 500 watts with a vacuum in V below 5×10^{-4} mm. of mercury. There is no reason why the above type of machine cannot be applied to any centrifuging problem, and it is our opinion that it is much superior to the air drive for the majority of cases.

Finally, attention should be called to another type of centrifuge which spins in a very high vacuum and is suspended by an axial magnetic suspension. This suspension was invented by F. T.

Holmes and has been developed at Virginia by Holmes, Smith, and MacHattie. The centrifuge is attached to a vertical iron rod which is suspended by the coaxial field of one or more solenoids. A steady direct current in one of the solenoids is not quite sufficient to support the rotating members, while the current through the other solenoid is controlled by a photoelectric cell arrangement. The amount of light striking the cell is increased as the centrifuge moves downward and decreased as it moves upward, making it possible to so adjust the current in the solenoid that the rod is automatically maintained at a predetermined height. Since the magnetic field is symmetrical around the vertical axis of the rod, there is no appreciable electromagnetic drag when the rod rotates. Consequently, when the rotor is mounted in vacuo there is almost a vanishing frictional torque against axial rotation. The "centrifuge" has been spun by rotating magnetic as well as electrostatic fields. These machines are especially suited to problems requiring extremely constant high rotational speed or to centrifuging at a reduced temperature or in a high vacuum.[3]

APPLICATIONS

The applications of high-speed centrifuges are so wide and varied that it will be impossible to discuss them here in detail. However, it might be of interest to mention some of the researches in which they have been used. Perhaps their widest and most important use up to the present time has been in the fields of medicine and biology. A great many biological and medical substances have been purified and their molecular weights and sizes studied. Svedberg and his collaborators, as well as Bauer, Pickels, Wyckoff, McFarland, Williams, Stern, Dietz, Severinghaus, Chiles, and others, have shown that most of the easily soluble native proteins are composed of homogeneous molecular species. Also they have found a regularity in the molecular weights of these proteins, which Svedberg interprets as indicating that their molecules are made up of definite fundamental

3. Recently MacHattie spun a .1 in. steel ball 6½ million r.p.m.

units which have masses of about 17,600. He assumes that only a few of these aggregates are stable, and the greater the molecular weight the fewer are the possibilities of stable aggregation. Some of the proteins have their structure altered by such things as dilution, change in pH value, or small amounts of foreign matter. For example, if the pH value of haemocyanin [4] of molecular weight 674×10^1 is changed by a small amount, it dissociates by steps into halves, eighths, and sixteenths, each dissociation product being a homogeneous molecular species. When the pH is changed back to the original value, the components recombine to form the original compound of molecular weight, 674×10^4. The homogeneity of the molecular species is determined by the sharpness of the sedimenting boundary in the ultracentrifuge, and within the limits of the resolving power of the ultracentrifuge most of the easily soluble proteins give sharp sedimenting boundaries. Substances such as tobacco mosaic virus and vaccinia virus, studied by Wyckoff and his collaborators, which have molecular weights well in excess of fifty millions, present important theoretical problems in molecular stability. Interesting differences in the number of sedimenting boundaries, as well as the relative concentration in each, between normal and pathological sera have been found. It is possible that the centrifuge may be very useful as a means of determining the correct diagnosis of disease.

In addition to proteins, many other substances have been studied by the centrifuge. For example, Lamm has shown that the masses of the molecules of starch depend upon its previous treatment and history and are, in general, polydisperse.[5] Signer has investigated Polystyrenes in different organic solvents and found results which indicate that in these solutions neighboring molecules greatly affect each other's free movement, except in very dilute solution. Nichols and Kraemer and others have

4. Respiratory pigment of the snail *Helix promatia* and other mollusks.

5. Recently Beckmann and his students have made a more detailed study of starch and arrived at a somewhat different conclusion.

studied a number of materials used industrially, such as rubber, Neoprene, cellulose, and cellulose derivatives, and have found relationships between the shape and weight of molecules and their macroscopic properties such as viscosity, etc. They have found that a great many of the synthetic large molecular weight substances are polydisperse. The particle sizes of both organic and inorganic colloids have been determined by centrifuging. Also it has been shown that the ultracentrifuge can distinguish between a simple mixture of molecules and chemical equilibrium. Furthermore, it gives the relationship between the equilibrium constant, pressure, temperature, and centrifugal field.

The electrical potentials developed in electrolytes when subjected to gravity or centrifugal fields were studied many years ago by Colley, Des Coudres, Burton, and Tolman. In the ultracentrifuge these potentials may reach values of the order of .1 volt when some electrolytes are first subjected to the centrifugal field. However, as the centrifuging is continued, the potential between the center and periphery decreases and theoretically, at least, should disappear completely when equilibrium is established. Comparing the molecular weight of the electrolyte obtained by the equilibrium method of centrifuging with the known value obtained from chemical data, it is possible to determine the activity coefficient or amount of dissociation of the electrolyte. Pedersen and Drucker have determined a number of activity coefficients in this way. In our own laboratory we have recently been studying both the electrical potentials and the concentration gradients set up in electrolytes in the ultracentrifuge. It might be noted in passing that with this type of technique it may be possible to identify the masses of artificially radioactive isotopes of an element.

As mentioned earlier, the high-speed ultracentrifuge has been used for the study of the relative densities of the materials in the living cell. Such things as "viscosity" and "surface tension" of the living substance, protoplasm, as well as the elasticity of the cell membranes, have been estimated. Also, it is helpful in dis-

tinguishing between real structures and so-called artifaxes. In some cases, after the cell has been centrifuged, abnormal development occurs, whereas in others this is not the case. The magnitude of the centrifugal fields required to kill some cells is rather surprising. H. W. Beams and R. L. King centrifuged fertilized eggs of the parasitic worm *Ascaris suum* at 400,000 gravity for an hour. The eggs were then removed from the centrifuge and observed under a microscope. At first the cell contents were stratified into definite layers. Twelve hours later the eggs had lost their stratification and most of them lived. However, the above treatment will kill a great many cells.

The centrifuge is of use in studies of the adhesion of films to surfaces. The experiments consist in determining the centrifugal force necessary to throw the film off the periphery of the centrifuge. Since the peripheral velocity of some of the air-driven, vacuum-type centrifuges may exceed twice the velocity of sound in air, interesting studies of the friction of different gases at reduced pressures on rapidly moving surfaces are possible. Several other applications could be discussed if space permitted, but perhaps the material presented will illustrate the extensive and important fields for the application of high-speed centrifuging technic.

X

RECENT ADVANCES IN AERONAUTICS

By J. C. Hunsaker

Massachusetts Institute of Technology

Advances in science become evident by their applications. In particular, the application of newly acquired knowledge in aerodynamics has produced the modern airplane, the latest child of our machine age. In its adolescence it seems to be a problem child and a precocious one. Its early promise of opening geographical barriers to free intercourse was once looked upon as furthering trade and cultural interchange and good understanding between nations. More recently, however, man has employed his new freedom in the third dimension to pursue old tribal quarrels and to threaten new ones.

A new understanding of physical phenomena leads to practical applications, but the uses to which we put them are unpredictable, though conditioned by the social organization and ethical standards of the time, and especially by the pressures under which economic life is conducted.

For example, an advance in organic chemistry may be applied to poison gas in time of war or to dye stuffs and medicines in time of peace. Printing records either trash or great literature. The radio carries music or jazz, education or propaganda.

We hope that the wise use of aviation will ultimately predominate, but in the meantime this country must vigorously pursue the application of aviation to our national defense, if we are to insure our continued existence as a free nation in the kind of world we live in.

The aspects of aviation which concern us immediately seem to be three;

First, as air power—to be used to enforce the national will.

Second, as a vehicle of air commerce, foreign and domestic.

Third, as a private vehicle of the air.

The present state of affairs, dependent on the relative expenditures of brains, money, and energy in each of these three aspects, indicates that the greatest development has taken place in air power, and the least in the service of the private owner. Considering the importance of the immediate results to be obtained, we need not be surprised that this is so.

Air Power

To appreciate the potentialities of air power is to understand why a disarmed and prostrate Germany seized upon this quickly forged weapon to make itself again the master of Europe. It seems likely that this weapon would have been seized by whatever national leader the German nation produced. In the years between the Treaty of Versailles and Hitler's rise to power, German aeronautical science made continuous progress in spite of and in some cases because of, the restrictions imposed upon it. Research and invention were directed toward economy and efficiency.

Under Hitler, the growth of the German air arm was an open secret. The disclosure of its potentialities dramatically readjusted the balance of power in Europe. The German nation was found to be in possession of air power greatly superior to that of its western neighbors. The fact that the Germans threatened to use it in September, 1938, marks a historical event of the first importance.

It seems unnecessary to dwell on the necessity for us to have an adequate national defense, including increase of air power. Advances in flying are rapidly reducing our geographical isolation. While the Atlantic and Pacific oceans still are broad enough to protect us against air raids started from Europe and Asia, we must not overlook the possibility of the establishment of hostile air bases in South America, the West Indies, or even in Mexico.

The islands of the Pacific are not only steppingstones for our own air mail route to Asia. With loss of control of the sea, these same islands could equally well serve an invader.

Guam, Hawaii, and Alaska, while of small concern to a true isolationist, can be of immense significance to the strategy of a foreign aggressor. The best insurance against a surprise, like that for Britain of September, 1938, is to be ourselves in the forefront of aviation development.

Air Commerce

Our domestic air lines lead the world in speed, safety, and volume of traffic and the trend is still upward [1]*. There have always been three difficulties:

The first is government regulation, which in the past has meant interference, but has now apparently been put on a sound basis by the recent creation of the Civil Aeronautics Authority [2].

The second is the weather. Ice formation [3] and blind landings [4] still constitute a serious hazard, but progress is coping with these conditions is encouraging and may be expected to be continuous.

The third difficulty is an inherent defect of the original invention of the Wright Brothers. Without adequate flying speed, control is lost and the airplane stalls. The airplane requires a high-speed take-off and landing as well as the maintenance of engine power in flight. Better airports, multiengined airplanes, and improved aerodynamic design are making for greater safety, but the ultimate solution is not in sight.

Private Flying

There is not much to be said about private flying, except that it has lagged. The small airplanes available to the private owner are costly to own and operate, demand a high degree of skill to fly, and are somewhat restricted in usefulness to the availability of landing fields and airports. Part of the difficulty lies in the

* Sources will be found in the list of References, p. 301.

inherent tendency of all airplanes to stall. The private owner, who is not a professional pilot, is more apt to get in trouble. Furthermore, elaborate radio aids and automatic devices to improve the safety of flight are to a large degree impractical for the small airplane. The accident rate is many times greater for miscellaneous flyers than for the common carriers.

Those who have imagined that the private plane market will grow like that for the private automobile have been disappointed. The rapid growth of the past ten years has been largely in the field of the common carrier or airline, with the sportsman coming last.

The automobile developed first with the sportsman, then with taxi and hire services, with the common carrier coming last. For the airplane, the steps in the process are just in reverse. But regardless of the order, it can be predicted with confidence that every year more people will fly for business and pleasure and that the next generation will take flying as a matter of course.

Recent Progress

The modern airplane is the result of recent advances in the aeronautical sciences applied to the Wright Brothers' original airplane. In 1908 this was a 40-mile biplane able to fly, with two men, for barely an hour [5]. Today the pursuit airplane exceeds 400 miles per hour in speed, bombers can fly at 300 miles, and transport planes cruise easily at 200 miles per hour [6]. These advances in performance have followed, somewhat discontinuously, new knowledge in aerodynamics, metallurgy, structural design, fuel technology, and engine and propeller design. The steps are sometimes abrupt as inventions occur, such as the variable pitch propeller or high-lift flaps. While the basic Wright invention still describes the present-day airplane, little remains of the original Wright design except its principle of control, which, as I said earlier, is its greatest defect, giving a tendency to stall unless a minimum flying speed is maintained.

Progress in the aeronautical sciences has been international in

its sources. For example, the basic concept of aerodynamic lift as a consequence of circulation is due to Lanchester [7] [11] of England, Kutta [8] [11] of Germany, and Joukowski [9] [11] of Russia and, in its final form, to Prandtl [10] [11] of Germany. Lord Rayleigh [12] introduced powerful methods of dimensional analysis in aerodynamics, and Bryan [13] and Bairstow [14], also of England, applied the theory of small oscillations to airplane stability.

From France came the braced monoplane, the fuselage, the single control stick, the wheeled landing gear, Drzewiecki's propeller theory [15], and Eiffel's wind-tunnel technique [16].

Prandtl in Germany has made perhaps the greatest single contribution to aerodynamic science with his theory of the boundary layer [17], by which classical hydrodynamics became practically useful. Wilm of Germany invented the strong light alloy, duralumin, which made possible both the Zeppelin and the modern Douglas monoplane [18].

In the field of design and invention there came from pre-Hitler Germany the Diesel engine and the Junkers cantilever wing [19]. Holland is represented by Fokker's stressed skin wing construction.

Practical advances from England, besides the first inherently stable airplane, include the Fairey flap [20] and Handley-Page slots [21] to increase lift, the air-cooled engine [22], and the sleeve-valve engine [23].

In America, fundamental scientific advances were largely imported, and the genius of our people shows best in design. The original Wright airplane was quickly improved by Glenn Curtiss with ailerons, to be followed by his first flying boat [24]. More recently, Reed's metal propeller [25] led to Caldwell's controllable pitch propeller [26], now run with a constant speed governor and arranged to be feathered when the engine stops [27]. Midgeley's research on knocking [28] has made high octane fuel universally available and resulted in engines of greater power and economy.

American achievements include the launching catapult [29], night flying, lighted airways, teletype for weather reporting, radio-range beacons [30], Mead's high performance air-cooled engines, the absolute altimeter [31], the Sperry gyroscopic compass and automatic pilot [32], retracting landing gear, "Alclad" [33] and stainless steel wings [34], sleeper planes, and a host of practical developments.

The National Advisory Committee for Aeronautics, by means of its wind-tunnel research at Langley Field, has contributed enormously to recent progress both at home and abroad. A partial list of contributions of Dr. George W. Lewis' organization should perhaps begin with the so-called N.A.C.A. cowling for the radial air-cooled engine, which cut in half the drag of the engine installation [35]. It is estimated that the saving to American airlines alone, from this feature, amounts to $2,500,000 every year.

The N.A.C.A. has also established the aerodynamic constants for a great series of wing profiles, to serve as a guide for designers [36]; and it has issued similar comprehensive information dealing with propeller design [37] and other basic features. It may truly be said that the Langley Field research program of the past twenty years has placed airplane design on a sound engineering basis.

The recently leading position of American aircraft is in large part to be attributed to Langley Field information. Naturally this information, when applied by our foreign competitors, produces equally beneficial results. But American designers have been more prompt to apply the results of research or perhaps they have had greater confidence in the importance and trustworthiness of the N.A.C.A. reports.

However, with the armament race on in Europe, there has been an enormous increase in foreign research efforts along the lines proved by us to be so productive, and today it is probable that American aircraft are no longer the best in the world. We may still be in the lead with commercial airplanes and flying boats, and with naval aircraft, but there is reason to believe that

the latest types of fighting and bombing airplanes produced in Europe fly faster and higher [38].

Technical advances in aviation take place when clever designers make use of new scientific knowledge. At any one time, there is no lack of good ideas for improvements but the good ideas are usually impractical. A technical advance takes place only when the time is ripe. High-compression engines were not practical until high octane fuel was available. Landing gears could not be retracted until thick cantilever wings were in use, and it was not worth while to retract them until the speed of flight became great enough to put a premium on saving the drag of such exposed parts in spite of increased weight and cost.

Aerodynamic Limitations on Progress

The basic knowledge that makes the airplane practical lies in the field of aerodynamics, and advances in aerodynamics set the trend of design and stimulate the adoption of nonaerodynamic features, which in themselves may lead to substantial improvements in performance.

For example, a better knowledge of the aerodynamics of skin friction indicated that rivet heads and minor roughness on wing surfaces increased the drag materially [39]. When speeds are high enough, this effect can be serious. We now see a recent trend toward very smooth polished surfaces for all high-speed aircraft.

Fundamentally, an airplane flies both because of and in spite of the air pressures on its surfaces. The weight is supported by the dynamic pressure of the air on the wings, caused by motion through the air. The thrust of the propeller maintains flying speed in spite of the drag of the air through which these same surfaces are moved.

The lift is the resultant vertical component of all surface pressures. The drag is made up of the resultant horizontal component of these same normal pressures, plus the frictional or shearing forces on the exposed surfaces. Consequently, the entire force of the air on the airplane is transmitted to its exposed surface

through the air in immediate contact with it, either as pressure or shear. If we could know the pressure and the shearing force at all points in the boundary layer of air next the surfaces, we could compute the resultant aerodynamic forces. Thus the distribution of pressure and shear in the boundary layer contains the key to lift and drag and to improvements in their ratio.

The solution of the equations of motion of a real fluid, having both inertia and density, is in general impossible. However, important technical progress has resulted from Prandtl's concept of a boundary layer [40], in which the motion is largely governed by viscous (shear) forces, surrounded by a streamline flow in which viscosity effects may be neglected. This simplifying assumption is justified by experiment, but we can readily see why it should be valid.

Shearing force due to viscosity depends directly on the relative amount of sliding or on the rate of change of velocity as we leave a surface. The air in contact with the surface of a solid sticks to it. There is no sliding, yet a short distance away the air is flowing with the velocity of the general stream. There is, therefore, a steep rise of velocity in passing from the surface through the boundary layer and a large shearing force is developed.

Outside the boundary layer, the streamlines of the general flow make a pattern in which changes in velocity from one streamline to the next are relatively small. In consequence, shearing forces there must be small. Prandtl's approximation makes them negligible.

On the other hand, since the time average of the flow in the boundary layer is essentially parallel to the surface, the pressure cannot change across the boundary layer. The pressure at any point on the surface is, therefore, that of the general flow immediately opposite, and is transmitted through the boundary layer substantially without change. We may have a pressure change or gradient along the surface and along the boundary layer, but not across it.

Classical hydromechanics can solve the equations of motion

for the flow of a frictionless fluid [41], and hence we can compute the pressure distribution over a wing surface so long as this pressure is transmitted through a boundary layer.

Likewise, we can compute the shearing forces on a surface from a knowledge of the viscosity of the air and the velocity distribution in the boundary layer [40].

Both methods of analysis fail completely, however, if the boundary layer separates from the surface and streams away as a layer of discontinuity, as imagined by Helmholtz and Kirchoff [42]. Obviously, Prandtl's physical concept of a boundary layer transmitting pressures to a boundary surface from an ideal continuous flow without no longer applies. Such flow separation takes place when a wing stalls [43], and is marked by an abnormal increase in drag. The practical prevention of separation and stalling is of vital importance in airplane design. The trigger which starts flow separation is evidently in the boundary layer, and future technical improvements seem to depend on a fundamental knowledge of boundary-layer mechanics. Presumably if we can control the boundary layer we can prevent separation.

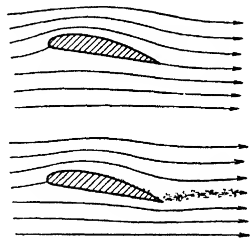
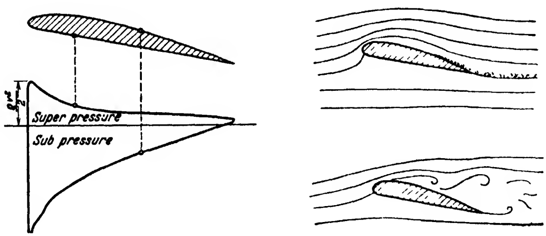


FIG. 117. Diagrams showing streamline pattern of flow of a frictionless fluid (above); and flow with friction (below).



a b

FIG. 118. (a) Pressure distribution of wing profile; (b) flow about a wing without and with separation.

Figure 117 shows diagrammatically, above, the streamline pattern of flow of a frictionless fluid, and below, the Prandtl concept of a boundary layer of viscous fluid surrounded by the ideal flow. Figure 118 shows the pressure distribution over a wing (a), giving rise to lift, and (b) the flow pattern before and after flow separation takes place.

Our present knowledge of the boundary layer is in accordance

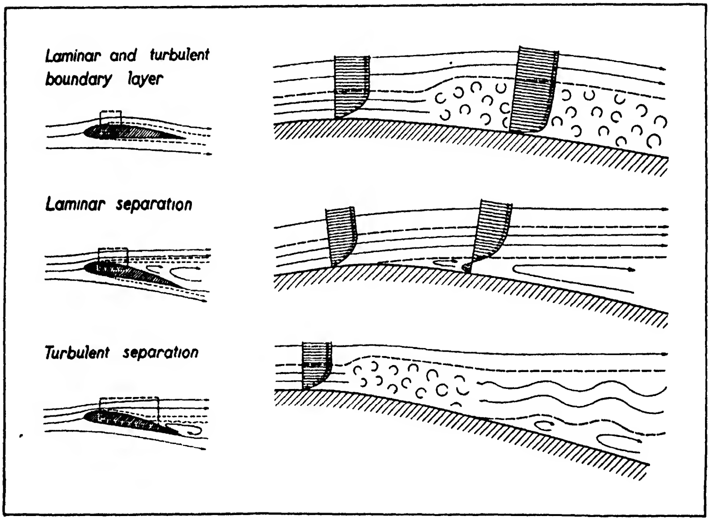


FIG. 119. Illustrating the concept of boundary-layer separation.

with the concept of the diagrams of Figure 119 [40]. The flow is at first laminar with a velocity gradient as shown. Then the flow reaches a transition point and becomes turbulent with a much steeper velocity gradient and increased frictional or shear force. However, if the surface be too much inclined or too strongly curved, the laminar boundary layer or the turbulent layer separates as shown, giving rise to a broad wake and large losses of energy. The flow in the general stream is distorted.

Transition to turbulent flow occurs earlier on a rough surface

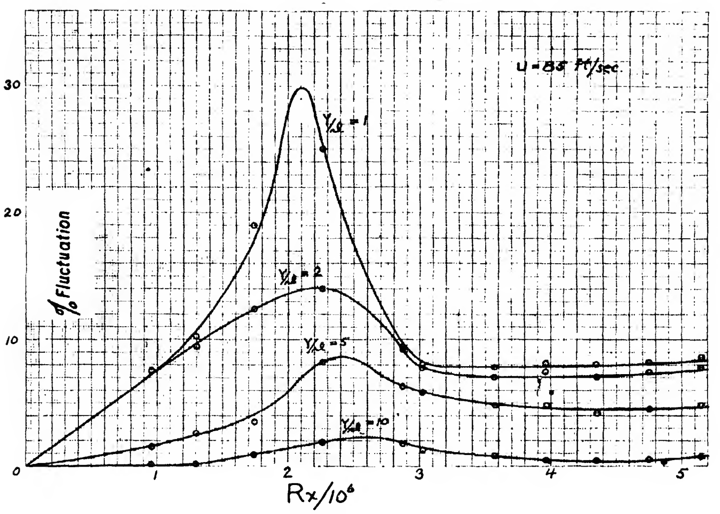


FIG. 120. Velocity fluctuations at given distances y from wall in terms of distance from leading edge (Rx). (After Peters.)

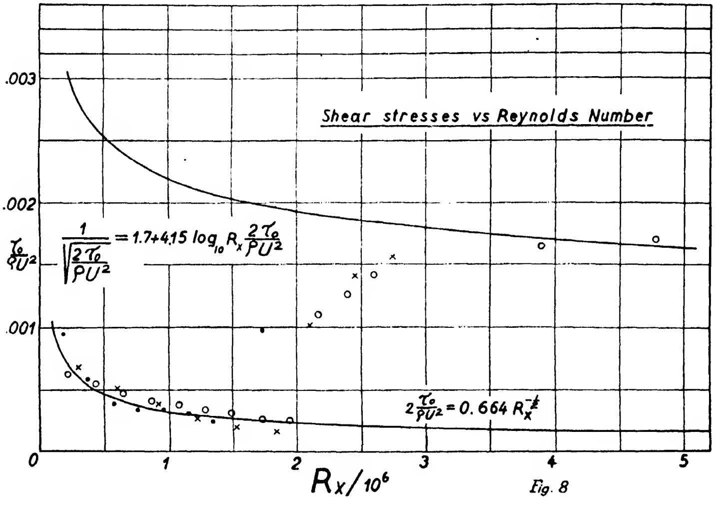


FIG. 121. Comparison of experimental values of shear stress with theoretical laminar and turbulent boundary-layer formulae. (After Peters.)

or for higher speeds. There is reason to believe that flow separation is marked by a thickening of the boundary layer and the accumulation of dead air and a reversal of velocity near the surface.

So long as the pressure along the top of the wing falls in the downstream direction, the boundary layer is kept in motion. When there is a rise of pressure, the flow is checked. If the rate of pressure rise be too great [43], flow separation occurs and the wing stalls. This phenomenon takes place sometimes suddenly over the wing tips with dangerous effects, including loss of aileron control.

In Figures 120 and 121, furnished by my colleague, H. Peters, there is recorded the percentage fluctuation of velocity measured with a hot-wire anemometer in the boundary layer of air flowing along a smooth flat plate. The point of transition from laminar to turbulent flow is clearly marked by a peak in velocity fluctuations with respect to the steady velocity of the general stream. The shearing force per unit area of surface, based on the measured velocity distributions across the boundary layer, is greatly increased with the onset of turbulent flow.

Figure 122 shows an extreme example of a stall where the separation takes place at the nose of the wing [43]. This is a photograph taken by the National Advisory Committee for Aeronautics using smoke to mark the streamlines. The laminar layer leaves the wing near the nose and becomes turbulent farther back. It is in effect a boundary layer of air between the general overrunning flow and the dead air on the back of the wing. Obviously, the drag of such a wing will be very high, due to loss of energy in a broad eddying wake stream. Likewise, the lift will be reduced since the suction over the top of the wing is destroyed when the flow separates.

Figure 123 shows another stalled wing, in this case provided with a trailing edge flap intended to increase its lift. The flap is without value when the flow does not follow the wing contour.

Figure 124 shows the same wing provided with slots on its back through which the boundary layer is sucked away. Separation

does not take place, and the general flow is seen to be strongly deflected downward by the wing and flap, thus creating a high lifting force.

The lift of a wing can be doubled by the use of a flap, but will stall as shown in Figure 125 at an inclination above 14°. By using boundary-layer control by suction, the lift of a plain wing can also be doubled. By the use of flaps and suction together the lift of a plain wing can be trebled.

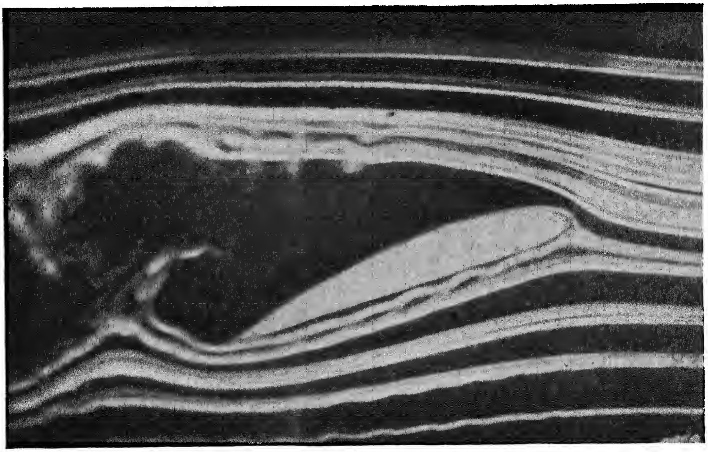


FIG. 122. Flow separation over top of airfoil. (N.A.C.A.)

Many such "high-lift devices" are being experimented with [44]. They all seek to prevent or delay flow separation and stalling. Figure 126 shows a large number of them, including the Handley-Page slots, which effectively speed up the boundary layer by leading a high-velocity jet of air into it from below. The propensity to stall, due to flow separation, as at the start and when landing, is a present limitation on high lift per unit wing area.

At high speed, the wings of an airplane are inclined so little, and bodies are made of such easy form, that flow separation does

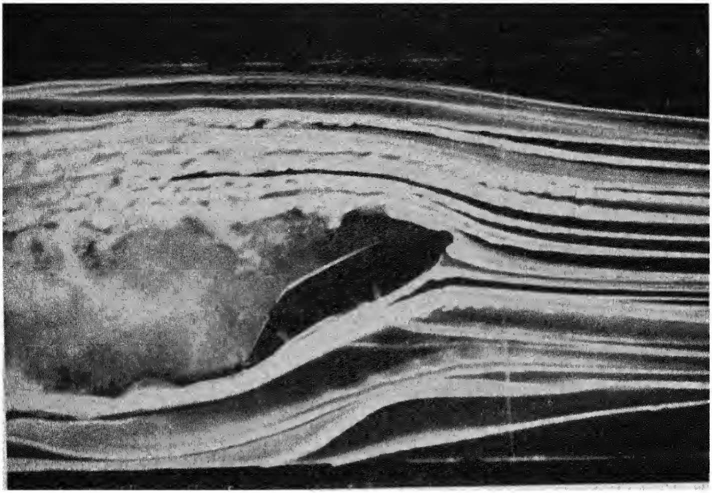


Fig. 123. Flow separation over airfoil with flap. (N.A.C.A.)

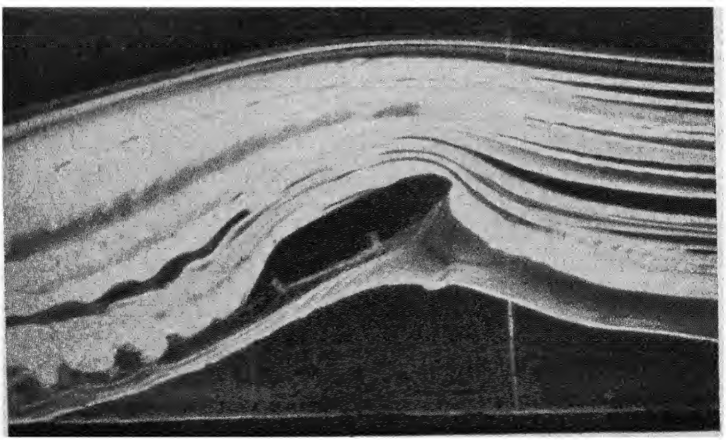


Fig. 124. Continuous flow over airfoil with flap. (N.A.C.A.)

not occur. The limitations on speed, therefore, appear to be only two: roughness and compressibility.

Roughness

The effect of roughness in increasing drag at high speed is due mainly to an earlier transition of the boundary layer from

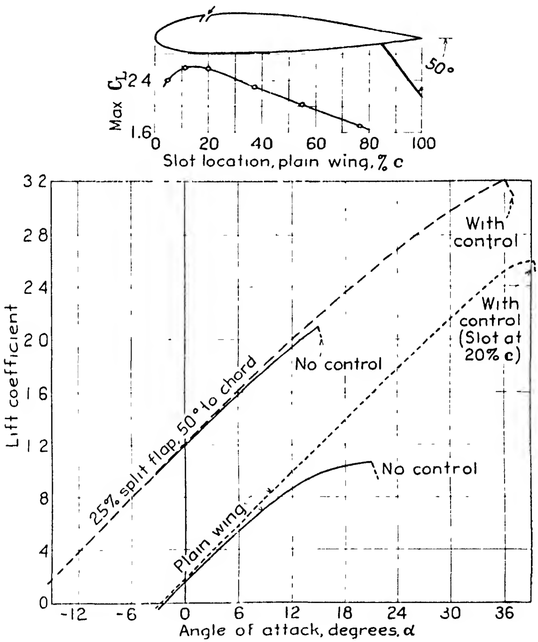


FIG. 125. Boundary-layer control on a sixteen-foot span. 2.67 foot chord. N.A.C.A. 2415 airfoil.

laminar to turbulent [401]. Presumably, due to the decrease of boundary-layer thickness with speed, a given degree of roughness is much more detrimental at high speed. Conversely, at some lower speed there is a permissible degree of roughness which is not detrimental.

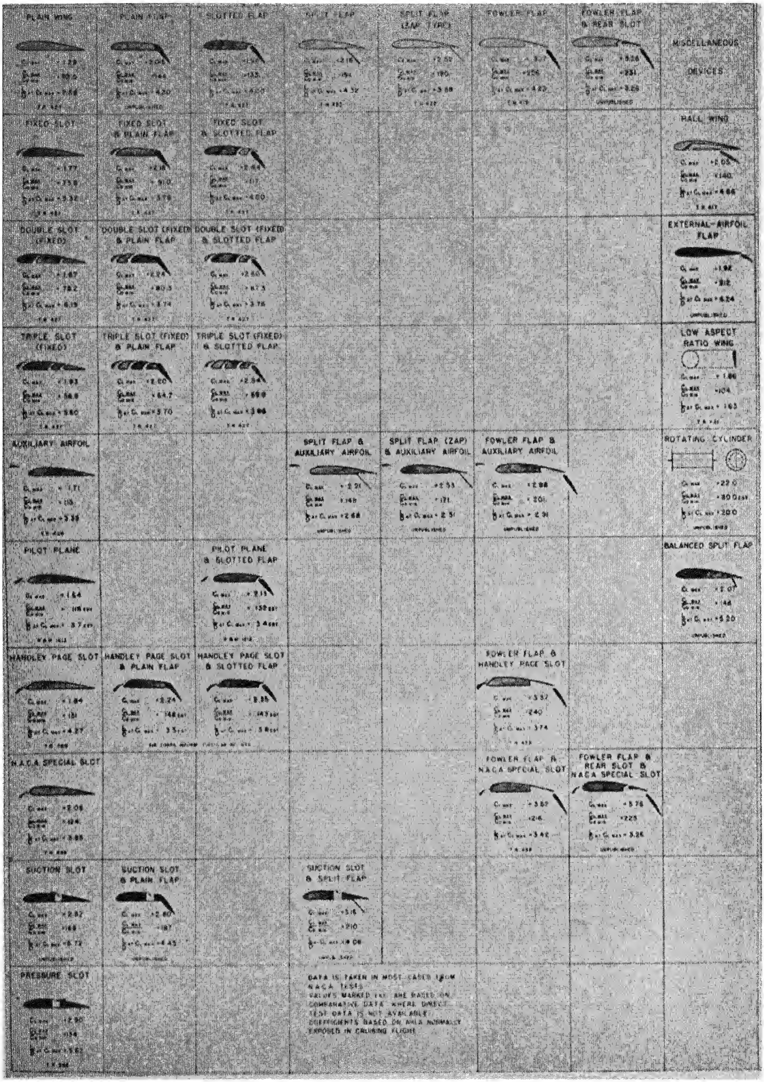


FIG. 126. Investigation of basic lift increasing devices and their combinations. Data are taken in most cases from N.A.C.A. tests. Values marked "est." are based on comparative data where direct test data are not available. Coefficients based on area normally exposed in cruising flight.

Figure 127 shows that a riveted wing of conventional service type has 42 per cent more drag than if it were truly smooth.

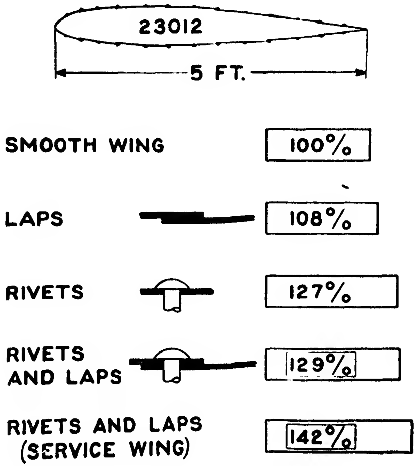


FIG. 127. Effects of rivets and laps at two hundred miles per hour. (N.A.C.A.)

Compressibility

As an experimental fact, it has long been known that airplane propellers are inefficient if the tip speed approaches the velocity of sound [45]. Confirmation of this fact is now found in a determination of the lift and drag coefficients at high speed of an airfoil such as is used both for propeller blades and wings. The high-speed wind tunnel of the National Advisory Committee for Aeronautics permits tests at extreme speeds. Figure 128, based on one of these tests, shows that, at some critical speed, the drag rises and the lift falls off abruptly. Note that for 10° incidence the drag begins to rise sharply above 250 miles per hour, while for — 2°, the drag holds constant up to 500 miles per hour. The world's speed record for airplanes is now held by Italy with a speed of

440 miles per hour [16]. A German fighting plane is reported to have exceeded 460 miles per hour [17].

It is evident that loss of effectiveness is a function of the compressibility of the air [18], which has been neglected in aviation until recently. The air at ordinary transportation speeds acts as if incompressible, but when the relative air velocity at any point over a wing reaches the speed of sound, we can readily imagine that a very different flow pattern must result.

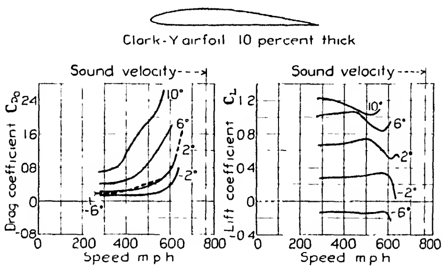


FIG. 128. Compressibility effect. (N.A.C.A.)

The aerodynamic force on a wing at ordinary transportation speeds, not near the velocity of sound, is created by the shear in the boundary layer and the normal pressures transmitted through the boundary layer from the general flow outside. However, for velocities exceeding the sound velocity, a new resistance is added; the so-called wave resistance or compressibility burble [49]. This new resistance is a pure pressure effect caused by a change in the pressure distribution over the surface of the wing as some critical speed is passed.

The explanation appears to lie in the entirely different flow pattern for supersonic velocities. For example, in an expanding nozzle, the velocity of a fluid decreases as the nozzle expands at ordinary velocities, but increases at supersonic velocities.

Likewise, for a converging nozzle, a fluid flow speeds up at ordinary speeds, but decelerates at supersonic speeds. Conse-

quently, any object which displaces the fluid acts somewhat like a converging nozzle. At supersonic speeds, the air flow over a thick wing, instead of speeding up and producing a strong suction, will, in fact, be slowed down. The resultant pressure increase accounts for increased drag and reduced lift.

The velocity of transmission of a pressure change in a con-

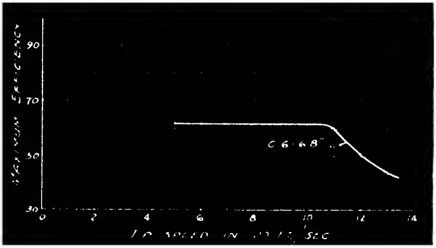


FIG. 129. Variation of propeller efficiency with tip speeds. (N.A.C.A.)

tinuous flow of fluid is the velocity of sound in that fluid. If the fluid be already moving at sound velocity, standing waves, called shock waves, form. Such shock waves are analogous to the standing waves on the surface of water when the water flows with a velocity greater than the velocity of propagation of the waves. Shock waves in the air flowing over a wing or projectile at supersonic speed can be photographed due to the change in density.

For wings of strong curvature, or inclination, it is entirely possible for the air to reach the velocity of sound at some place on the wing, while the general stream velocity is still much less than seven hundred and fifty miles per hour, the velocity of sound. This seems to be the explanation of Figure 128, in which the "compressibility burble" occurs at lower speeds for larger angles of inclination.

The idea suggests itself that if our airplanes are to fly faster than 400 miles per hour, we had better sharpen their forms, avoiding all strong curvature in the longitudinal direction. To

set a limit on the possible speed of airplanes, we would need to know where the compressibility burble would occur on other parts of the airplane, including the propeller. It is obvious, however, that there is a limit and that we are approaching it. Perhaps the propeller will not be the best means of propulsion for extreme flight velocities and some form of rocket propulsion may ultimately be sought.

CONCLUSION

We have considered some of the fundamental limitations to the future progress of airplanes so far as they are to be found in the field of aerodynamics. Surface smoothness is evidently of great importance for higher speed. Compressibility, however, appears to place a real limit on high speed, which may be raised a little as we learn more of high-speed boundary-layer mechanics and the cause of shock waves and sound radiation.

Greater economy of power at more moderate speeds may also be obtained by a better understanding of the boundary layer, and the role of surface finish and curvature in promoting the transition from laminar to turbulent flow.

Again the boundary layer appears to hold the key to flow separation and its dangerous attendant, stalling. If the present-day airplane leaves something to be desired, it is principally on the score of safety against stalling.

Research and experiment lead to invention and technical improvements. A consistent program of research into every aspect of the boundary layer would appear to be fundamental to future progress.

Any program of development, based on research, requires both coördination and decentralization. We have in this country effective coördination of research by the National Advisory Committee for Aeronautics, and a degree of decentralization in the practice of competitive bidding for the supply of aircraft for the army, navy, and airlines. As more independent scientists in the schools and universities become interested in the scientific prob-

lems of aviation, we may expect our democratic system to hold its position. Unpredictable inventions are certain to occur when many men think and experiment in a free competitive organization.

REFERENCES

CHAPTER I

GOODSPEED, T. H. The effects of x rays and radium on species of the genus Nicotiana. *Jour. Hered.* *20*, 243–259, 1929.

MCCLINTOCK, B. Cytological observations of deficiencies involving known genes, translocations and an inversion in Zea mays. Missouri Agr. Exp. Sta., Res. Bull. 163, 1931.

MORGAN, T. H. On the mechanism of heredity. Croonian Lecture. *Proc. Roy. Soc.* (London), *94*, 162–197, 1922.

MULLER, H. J. Artificial transmutation of the gene. *Science*, *66*, 84–87, 1927.

—— The problem of genic modification. *Verh. V int. Kong. Vererbungsw.*, pp. 234–260. *Z. indukt, Abstamm.-u. Vererblehre*, Suppl. I, 1928.

—— Further studies on the nature and causes of gene mutations. *Proc. Sixth Int. Cong. Genetics* (Ithaca), *1*, 213–255, 1932.

—— and MACKENZIE, K. Discriminatory effect of ultraviolet rays on mutations. *Nature* (London), *143*, 83–84, 1939.

PAINTER, T. S. A new method for the study of chromosome aberrations and the plotting of chromosome maps in *Drosophila melanogaster*. *Genetics*, *19*, 175–188, 1934.

RHOADES, M. M., and MCCLINTOCK, B. The cytogenetics of maize. *Bot. Rev.* *1*, 292–325, 1935.

SAX, K. Chromosome aberrations induced by X rays. *Genetics*, *23*, 494–516, 1938.

SCHULTZ, J. Radiation and the study of mutation in animals. Biological Effects of Radiation, 1209–1261, 1936. Edited by B. M. Duggar.

STADLER, L. J. Genetic effects of X rays in maize. *Proc. Nat. Acad. Sci.*, *14*, 69–75, 1938.

—— On the genetic nature of induced mutations in plants. *Proc. Sixth Int. Cong. Genetics* (Ithaca), *1*, 274–294, 1932.

—— and SPRAGUE, G. F. Genetic effects of ultraviolet radiation in maize. I. Unfiltrated radiation. II. Filtered radiations. III. Effects of nearly monochromatic λ2537, and comparison of effects of X-ray and ultraviolet treatment. *Proc. Nat. Acad. Sci.* (Wash.), *22*, 572–578, 579–583, 584–591, 1936.

STADLER, L. J. Induced Mutations in Plants. Biological Effects of Radiation, 1209–1261, 1936. Edited by B. M. Duggar.

CHAPTER II

1. For a more complete review see Went and Thimann, Phytohormones, New York, 1937.
2. ROBBINS, W. J. Effect of autolized yeast and peptone on growth of excised corn root tips in the dark. *Bot. Gaz.*, *74*, 59–79, 1922.
3. WHITE, P. R. Potentially unlimited growth of excised tomato root tips in a liquid medium. *Plant Physiol.*, *9*, 585–600, 1934.
4. BONNER, J. Vitamin B_1 a growth factor for higher plants. *Science*, *85*, 183–184, 1937.
5. ROBBINS, W. J., BARTLEY, M. A. (1) Vitamin B_1 and the growth of excised tomato roots. *Science*, *85*, 246–247, 1937.
6. ADDICOTT, F. T., and BONNER, J. Nicotinic acid and the growth of isolated pea roots. *Science*, *88*, 577–578, 1938.
7. ROBBINS, W. J., and SCHMIDT, M. B. Further experiments on excised tomato roots. *Am. Jour. Bot.*, *26*, 149–159, 1939.
8. BOYSEN-JENSEN, P. La transmission de l'irritation phototropique dans l'Avena. *Bull. Acad. Roy. Denmark*, *1*, 3–24, 1911.
9. PAAL, A. Ueber phototropische Reizleitung. *Jahrb. wiss. Bot.*, *58*, 406–458, 1919.
10. WENT, F. W. Wuchsstoff und Wachstum. *Rec. trav. bot. neerl.*, *25*, 1–116, 1928.
11. KOGL, F., and HAAGEN-SMIT, A. J. Über die Chemie des Wuchsstoffs. *Proc. Kon. Akad. Amsterdam*, *34*, 1411–1416, 1931.
12. KOGL, F., and HAAGEN-SMIT, A. J., and ERXLEBEN, H. Über ein neues Auxin ("Hetero-auxin") aus Harn. *Z. physiol. Chem.*, 228, 90–103, 1934.
13. DOLK, H. E. Geotropie en groeistof. Diss. Utrecht, 1930.
14. WENT, F. W. On a substance causing root formation. *Proc. Kon. Acad. Wetensch. Amsterdam*, *32*, 35–39, 1929.
15. THIMANN, K. V., and KOEPFLI, J. B. Identity of the growth-promoting and root-forming substances of plants. *Nature*, *135*, 101, 1935.
16. WENT, F. W., BONNER, J., and WARNER, G. C. Aneurin and the rooting of cuttings. *Science*, *87*, 170–171, 1938.
17. GUSTAFSON, F. G. Inducement of fruit development by growth promoting chemicals. *Proc. Nat. Acad. Sci.*, *22*, 628–636, 1936.
18. GARDNER, F. E., MARTH, P. C., and BATJER, L. P. Spraying with plant

growth substances to prevent apple fruit dropping, *Science*, *90*, 208–209, 1939.
19. THIMANN, K. V. On an analysis of the activity of two growth-promoting substances on plant tissues. *Proc. Kon. Akad. Wetensch. Amsterdam*, *38*, 896–912, 1935.
20. BONNER, D. M. Relation of environment and of the physical properties of synthetic growth substances to the growth reaction. *Bot. Gaz.*, *100*, 200–214, 1938.
21. KOEPFLI, J. B., THIMANN, K. V., and WENT, F. W. Phytohormones: Structure and physiological activity, I. *Jour. Biol. Chem.*, *122*, 763–780, 1938.
22. OVERBEEK, J. VAN. The growth hormone and the dwarf type of growth in corn. *Proc. Nat. Acad. Sci.*, *21*, 292–299, 1935.
23. BONNER, D. M., and HAAGEN-SMIT, A. J. Leaf growth factors. II. The activity of pure substances in leaf growth. *Proc. Nat. Acad. Sci.*, *25*, 184–188, 1939.
24. HAMNER, K. C., and BONNER, J. Photoperiodism in relation to hormones as factors in floral initiation and development. *Bot. Gaz.*, *100*, 388–431, 1938.

CHAPTER III

BAILEY, P., and BREMER, F. A sensory cortical representation of the vagus nerve with a note on the effects of low blood pressure on the cortical electrogram. *Jour. Neurophysiol.*, *1*, 405–412, 1938.

—— and SWEET, W. H. Effects of respiration, blood pressure, and gastric motility of stimulation of the orbital surface of the frontal lobe. *Jour. Neurophysiol.*, *3*, 276–281, 1940.

BEEVOR, C. E., and HORSLEY, V. A record of the results obtained by electrical excitation of the so-called motor cortex and internal capsule in an orang-outang (*Simia satyrus*). *Philos. Trans.*, *181* B, 129–158, 1890.

BIEBER, I., and FULTON, J. F. The relation of the cerebral cortex to the grasp reflex and to the postural and righting reflexes. *Arch. Neurol. Psychiat.* (Chicago), *39*, 435–454, 1938.

BOERNSTEIN, W. S. The cortical taste area in monkeys and a semiquantitative method of testing taste in monkeys. *Jour. Am. Physiol.*, *129*, 314, 1940.

BRODMANN, K. Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues. J. A. Barth, Leipzig, 1909; reprinted, 1925, xii, 324 pp.

BUCY, P. C., and CASE, T. J. Cortical innervation of respiratory movements.

I. Slowing of respiratory movements by cerebral stimulation. *Jour. Nerv. Ment. Dis.*, *84*, 156–168, 1936.

Catchpole, H. R., and Fulton, J. F. Tarsiers in captivity. *Nature*, *144*, 514, 1939.

Clark, W. E. LeGros. Notes on the living tarsier (*Tarsius spectrum*). *Proc. Zool. Soc.* (London), 217–223, 1924.

—— Early forerunners of man. A morphological study of the evolutionary origin of the primates. Ballière, Tindall and Cox, London, 1934, xvi, 296 pp.

Cook, Norman. Notes on a group of tarsiers in captivity. *Jour. Mammal.*, *20*, 173–178, 1939.

Dusser de Barenne, J. G., and McCulloch, W. S. Functional organization in the sensory cortex of the monkey (*Macaca mulatta*). *Jour. Neurophysiol.*, *1*, 69–85, 1938.

—— McCulloch, W. S., and Ogawa, T. Functional organization in the face-subdivision of the sensory cortex of the monkey (*Macaca mulatta*). *Jour. Neurophysiol.*, *1*, 436–441, 1938.

Ferrier, D. The functions of the brain. Smith, Elder and Co., London, 1876, xv, 323 pp.

Finan, J. L. Effects of frontal lobe lesions on temporally organized behavior in monkeys. *Jour. Neurophysiol.*, *2*, 208–226, 1939.

Freeman, W., and Watts, J. W. Prefrontal lobotomy in the treatment of mental disorders. *Jour. Sth. Med. Ass.*, *30*, 23–31, 1937.

Fulton, J. F., and Dusser de Barenne, J. G. The representation of the tail in the motor cortex of primates, with special reference to spider monkeys. *Jour. Cell. Comp. Physiol.*, *2*, 399–426, 1933.

—— and Keller, A. D. The sign of Babinski. A study of the evolution of cortical dominance in primates. Charles C. Thomas, Springfield, Ill., 1932, xv, 165 pp.

——, Ranson, S. W., and Frantz, A. M. (*editors*). The hypothalamus and central levels of autonomic function. *Res. Publ. Ass. Nerv. Ment. Dis.*, vol. *20*, 1940, xxx, 980 pp.

Green, H. D., and Walker, A. E. The effects of ablation of the cortical motor face area in monkeys. *Jour. Neurophysiol.*, *1*, 262–280, 1938.

Grünbaum, A. S. F., and Sherrington, C. S. Observations on the physiology of the cerebral cortex of some of the higher apes. *Proc. Roy. Soc.*, *69*, 206–209, 1901.

Hilgard, E. R., and Marquis, D. G. Conditioning and learning. Appleton-Century Co., New York, 1940, xii, 429 pp.

Jacobsen, C. F. The effects of extirpations on higher brain processes. *Physiol. Rev.*, *19*, 303–322, 1939.

Kennard, Margaret A. A reorganization of motor function in the cerebral

cortex of monkeys deprived of motor and premotor areas in infancy. *Jour. Neurophysiol.*, *1*, 477–496, 1938.

KENNARD, MARGARET A. Alteration in response to visual stimuli following lesions of frontal lobe in monkeys. *Arch. Neurol. Psychiat.*, *41*, 1153–1165, 1939.

—— and ECTORS, L. Forced circling in monkeys following lesions of the frontal lobes. *Jour. Neurophysiol.*, *1*, 45–54, 1938.

LASHLEY, K. S. Studies of cerebral function in learning. XI. The behavior of the rat in latch box situations. The mechanism of vision. XII. Nervous structures concerned in the acquisition and retention of habits based on reactions to light. *Comp. Psychol. Mongr.*, No. 52, *11*, 1935, 79 pp.

LEYTON, A. S. F., and SHERRINGTON, C. S. Observations on the excitable cortex of the chimpanzee, orang-utan and gorilla. *Quart. Jour. Exp. Physiol.*, *11*, 135–222, 1917.

LINNÉ, C. VON. *Systema naturae, sive regna tria naturae systematica proposita per classes, ordines, genera, et species.* 1st ed., T. Haak, Leyden, 1735, 14 fol.

LYERLY, J. G. Prefrontal lobotomy in involutional melancholia. *Jour. Fla. Med. Ass.*, 1938.

—— Transsection [*sic*] of the deep association fibers of the prefrontal lobes in certain mental disorders. *Southern Surgeon*, *8*, 426–434, 1939.

MAGOUN, H. W. Excitability of the hypothalamus after degeneration of corticifugal connections from the frontal lobes. *Am. Jour. Physiol.*, *122*, 530–541, 1938.

MARQUIS, D. G. Effects of removal of the visual cortex in mammals, with observations on the retention of light discrimination in dogs. *Res. Publ. Ass. Nerv. Ment. Dis.*, *13*, 558–592, 1934.

—— Phylogenetic interpretation of the functions of the visual cortex. *Arch. Neurol. Psychiat.* (Chicago), *33*, 807–815, 1935.

—— and HILGARD, E. R. Conditioned responses to light in monkeys after removal of the occipital lobes. *Brain*, *60*, 1–12, 1937.

MATTHEW, W. D. The evolution of the mammals in the Eocene. *Proc. Zool. Soc.* (London), 1928, 947–985.

PITTS, R. F., and MAGOUN, H. W., and RANSON, S. W. Localization of the medullary respiratory centers in the cat. *Am. Jour. Physiol.*, *126*, 673–688, 1939.

—— MAGOUN, H. W., and RANSON, S. W. Interrelations of the respiratory centers in the cat. *Am. Jour. Physiol.*, *126*, 689–707, 1939.

RYLANDER, G. Personality changes after operation on the frontal lobes. A clinical study of 32 cases. Munksgaard, Copenhagen, 1939.

SCHÄFER, E. A. Experiments on special sense localizations in the cortex cerebri of the monkey. *Brain*, *10*, 352–380, 1888.

SCHÄFER, E. A. On the alleged sensory functions of the motor cortex cerebri. *Jour. Physiol.*, *23*, 310–314, 1898.

SCHRADER, M. E. G. Zur Physiologie des Vogelhirns. *Pflüg. Arch. ges. Physiol.*, *44*, 175–238, 1889.

SHERRINGTON, C. S. Notes on the arrangement of motor fibres in the lumbo-sacral plexus. *Jour. Physiol.*, *13*, 621–772, 1892.

—— Flexion-reflex of the limb, crossed extension-reflex, and reflex stepping and standing. *Jour. Physiol.*, *40*, 28–121, 1910.

—— Quantitative management of contraction in lowest level coördination. Hughlings Jackson Lecture. *Brain*, *54*, 1–28, 1931.

SMITH, W. K. The representation of respiratory movements in the cerebral cortex. *Jour. Neurophysiol.*, *1*, 55–68, 1938.

STOCK, CHESTER. An Eocene primate from California. *Proc. Nat. Acad. Sci.* (Wash.), *19*, 954–959, 1933.

—— A second Eocene primate from California. *Proc. Nat. Acad. Sci.* (Wash.), *20*, 150–154, 1934.

WALKER, A. E., and GREEN, H. D. Electrical excitability of the motor face area. A comparative study in primates. *Jour. Neurophysiol.*, *1*, 152–165, 1938.

WATTS, J. W., and FREEMAN, W. Psychosurgery. Effect on certain mental symptoms of surgical interruption of pathways in the frontal lobes. *Jour. Nerv. Ment. Dis.*, *88*, 589–601, 1938; see also *Yale Jour. Biol. and Med.*, *11*, 527–545, 1939.

WOODWARD, A. S., SMITH, G. ELLIOT, HILL, J. P., WOOD-JONES, F., POCOCK, R. I., CUNNINGHAM, J. T., MITCHELL, P. C., and MACBRIDE, E. W. Discussion on the zoological position and affinities of *Tarsius*. *Proc. Zool. Soc.* (London), *23*, 465–498, 1 pl., 1919.

WOOLLARD, H. H. The anatomy of *Tarsius spectrum*. *Proc. Zool. Soc.* (London), 1071–1184, 1935.

CHAPTER IV

COOKE, C. WYTHE. Discussion of the Origin of the Supposed Meteorite Scars of South Carolina. *Jour. Geol.*, *42*, 88–96, 1934.

—— Geology of the Coastal Plain of South Carolina. U. S. Geol. Surv., Bull. *867*, 1936, 196 pp.

GLENN, L. C. Some Notes on Darlington (S.C.), "Bays." *Science*, *2*, 472–475, 1895.

JOHNSON, DOUGLAS. Supposed Meteorite Scars of South Carolina. *Science*, *79*, 461, 1934.

Johnson, Douglas. Origin of the Carolina "Bays." *Proc. Geol. Soc. Am.*, 1935, *84*, 1936.

—— Origin of the Supposed Meteorite Scars of Carolina. *Science*, *84*, 15–18, 1936.

MacCarthy, Gerald R. The Carolina "Bays." *Proc. Geol. Soc. Am.*, 1935, 90–91, 1936.

Melton, F. A., and Schriever, William. Meteorite Scars in the Carolinas. Geol. Soc. Am., Bull. *44*, 94, 1933.

—— The Carolina "Bays"—Are They Meteorite Scars? *Jour. Geol.*, *41*, 52–66, 1933.

Melton, F. A. The Origin of the Carolina "Bays." *Discovery*, *15*, 151–154, 1934.

—— Reply. [to Cooke, C. Wythe. Discussion of the Origin of the Supposed Meteorite Scars of South Carolina.] *Jour. Geol.*, *42*, 97–104, 1934.

Nininger, H. H. Our Stone-Pelted Planet. Houghton Mifflin Co., Boston, 1933, 237 pp.

Prouty, William F. Carolina "Bays" and Elliptical Lake Basins. *Jour. Geol.*, *43*, 200–207, 1935.

—— Further Evidence in Regard to the Origin of Carolina "Bays" and Elliptical Lake Basins. *Proc. Geol. Soc. Am.*, 1935, 96–97, 1936.

Raisz, Erwin. Rounded Lakes and Lagoons of the Coastal Plains of Massachusetts. *Jour. Geol.*, *42*, 839–848, 1934.

Watson, Fletcher, Jr. Meteor Craters. *Pop. Astron.*, *44*, 2–17, 1936.

Wylie, C. C. On the Formation of Meteoric Craters. *Pop. Astron.*, *41*, 211–214, 1933.

CHAPTER V

Abbot, C. G. Solar Radiation and Weather Studies. Smithsonian Inst., Miscell. Coll., Vol. 94, No. 10, 1935.

Bradley, W. H. The Varves and Climate of the Green River Epoch. U. S. Geol. Surv., Professional Paper 158, 1929.

—— Non-Glacial Marine Varves. *Am. Jour. Sci.*, *22*, 321, 1931.

—— *Scientific Monthly*, May, p. 421, 1936.

—— The Biography of an Ancient American Lake. Ann. Report, Smithsonian Inst., pp. 279–290, 1937.

——Non-Glacial Varves, with Selected Bibliography. Report Comm. on Measurement of Geol. Time, May 1, 1927, pp. 32–43, 1937. See National Research Council.

Brewer, A. K. Isotopes of Potassium. *Ind. and Eng. Chem.*, *30*, 892–896, 1938.

BREWER, A. K. See also National Research Council Bibliography, Comm. Report, p. 15, 1939. (Division of Geology and Geography, Committee on Measurement of Geologic Time.)

BROWN, H. JANVRIN. Long Range Weather Service, No. 3, March 16, 1925. Copyright by Herbert Janvrin Brown, 1214 L. Street, Washington, D.C. (This material is rarely found in libraries. The data are largely to be found in works by E. W. Brown, E. Huntington, H. H. Clayton, etc. Brown had a lively war with the government weather service!)

CLAYTON, H. H. Centers of Action and Long Period Weather Changes. Am. Meteorological Soc., Bull. *19*, 27–28, 76–77, 1938.

COLBERT, H. F., and OSBORN, H. F. See Osborn.

COLLET, L. W. Les Lacs. Gaston Doin, Publisher, Paris, 1925.

DOLE, M. The Relative Atomic Weight of Oxygen in Water and in Air. *Jour. Chem. Physics*, *4*, 268–275, 1936.

ELLSWORTH, H. V. Rare Elements of Canada. Bull. Canada Dept. Mines, Econ. Ser., No. 11, 1932. Summarizes many papers in the *American Mineralogist* from 1926 on. Since then his work has been abstracted in the National Research Council Committee Reports and Bibliographies.

EVANS, R. D. The Determination of the Age of Iron and Stony Meteorites from Radioactivity. *Pop. Astronomy*, *46*, 1–12, 1938.

—— Numerous papers, often with other authors. See National Research Council Comm. Bibliography, pp. 21–22, 1939.

EVANS, R. D., GOODMAN, C., KEEVIL, N. B., LANE, A. C., and URRY, W. D. Intercalibration and Comparison in Two Laboratories of Measurements Incident to the Determination of the Geological "Age of Rocks." *Phys. Rev.*, *55*, 931–939, 1939.

FREEMAN, J. R. Regulation of Elevation and Discharge of the Great Lakes. A Report to the President and Board of the Sanitary District of Chicago. Includes letters from F. B. Taylor, A. C. Lane, etc. Published in Providence, 1926, probably by the author, 548 pp.

GUTENBERG, B., et al. Internal Constitution of the Earth. McGraw-Hill, 1939.

HOLMES, A. The Age of the Earth. T. Nelson and Sons, 1937. Rewritten from 1913 and 1927 editions.

—— See National Research Council, Bull. *80*.

JEFFREYS, H. The Earth. Macmillan Co., 1929.

—— See other papers indexed by Gutenberg and Holmes.

KORN, H. Schichtung und absolut Zeit. *N. Jahrb. Min. Geol.*, 74A, 50, 1938.

KNOPF, A. See National Research Council, Bull. *80*.

LANE, A. C. Rating the Geological Clock (1934). *Proc. Sixteenth Int. Geol. Congress* (Wash.), *II*, 146–167, 1936.

LANE, A. C. Measuring Geologic Time. Its Difficulties. Smithsonian Report for 1937, pp. 235–254.
—— See also National Research Council.
MARBLE, J. P. See National Research Council.
National Research Council. This organization has published a number of printed and mimeographed reports, generally by more than one author. In addition to the following, the Division of Geology and Geography has a Committee on the Measurement of Geologic Time of which, at present, A. C. Lane is chairman and J. P. Marble vice chairman, and whose annual Reports have been mimeographed and bound since 1932. A select bibliography by R. C. Wells was issued in 1924 together with the reports from 1932–35. Since then Marble has annually contributed an annotated bibliography and has appended special papers, and exhibits of various workers.
—— Bull. *51*. KOVARIK, A. F., and MCKEEGAN, L. W. Radioactivity, 1929.
—— Bull. *80*, 1931.
—— BROWN, E. W. The Age of the Earth from Astronomical Data.
—— HOLMES, A. Radioactivity and Geological Time.
—— KNOPF, A. Editor of "The Age of the Earth" in *Physics of the Earth*. Also author of summary and Part 2 on "The Age of the Ocean."
—— KOVARIK, A. F. Calculating the Age of Minerals from Radioactivity Data and Principles.
—— SCHUCHERT, C. Geochronology or The Age of the Earth on the Basis of Sediments and Life.
NIER, A. O. I. The Isotopic Composition of Uranium and the Half Lives of the Uranium Isotopes. II. The Isotopic Constitution of Radiogenic Lead and the Measurement of Geologic Time. *Phys. Rev.*, *55*, 150–163, 1939.
—— and GULBRANSEN, E. A. Variation in the Relative Abundance of the Carbon Isotopes. *Jour. Am. Chem. Soc.*, *61*, 697, 1939.
—— BAXTER, G. P., and GULBRANSEN, E. O. See Abstracts in National Research Council Comm. Reports.
OSBORN, H. F. Biological Inductions from the Evolution of the Proboscidea. *Science*, *76*, 503, 1932.
—— and COLBERT, E. H. The Elephant Enamel Method of Measuring Pleistocene Time. *Proc. Am. Phil. Soc.*, *70*, 187, 1931.
ROSE, J. L., and STRANATHAN, R. K. Geologic Time and Isotopic Composition of Radiogenic Lead. *Phys. Rev.*, *50*, 791–796, 1936.
RUBEY, W. W. Lithologic Studies of Fine Grained Upper Cretaceous Sedimentary Rocks of the Black Hills Region. U. S. Geol. Surv., Professional Paper 165, pp. 1–54, 1930.
SPENCE, H. S. The Pitchblende and Silver Discoveries at Great Bear Lake, Northwest Territories. Dept. of Mines, Canada, 1932.

TAYLOR, F. B. U. S. Geol. Surv., Atlas, Niagara Folio, No. 90, 1913.
—— The Status of Lake Erie in Present and Recent Land-Tilting. Papers of the Michigan Acad. of Sci., Arts and Letters, pp. 251–260, 1928.
—— See Letters in Report of J. R. Freeman, 1926.
UREY, H. C. The Separation of Isotopes and Their Use in Chemistry and Biology, chap. II, Science in Progress, edited by Baitsell, Yale University Press, New Haven, 1939.
URRY, W. D. Helium and the Problem of Geological Time. *Chem. Reviews*, *17*, 305–343, 1933.
—— Ages by the Helium Method, I. Geol. Soc. Am., Bull. *46*, 1101–1146, 1935.
—— 1939. See Evans, R. D., above. A paper correcting some previously given figures.
—— The reports of the Committee of the National Research Council, cited above, contain reports of Urry's activities from 1931 on.

CHAPTER VI

I. *General Nontechnical References.*

EDDINGTON, A. S. The Expanding Universe. Macmillan, New York, 1933. Popular account.
HUBBLE, E. The Realm of the Nebulae. Yale, 1936. Methods and results of nebular research.
—— The Observational Approach to Cosmology. Oxford, 1937. Shorter account, emphasizing number-magnitude relation.
JOHNSON, Francis R. Astronomical Thought in Renaissance England. Johns Hopkins, Baltimore, 1937. Chap. VI on Digges, Bruno, and the infinity of the universe.
—— and LARKEY, SANFORD V. Thomas Digges, the Copernican System, and the Idea of the Infinity of the Universe in 1576. *Huntington Library Bull.*, *5*, 69–117, 1934.
LIEBER, L. R., and H. G. Non-Euclidean Geometry, or Three Moons in Mathesis. New York, 1931. Delightful and charmingly illustrated nontechnical chat on non-Euclidean geometry.
POINCARÉ, H. The Foundations of Science. Science Press, New York, 1929. Poincaré's attitude toward geometry; in particular cf. chap. V of Science and Hypothesis.
SHAPLEY, H. Flights from Chaos. McGraw-Hill, New York, 1930. Subtitled "A survey of material systems from atoms to galaxies."
DE SITTER, W. Kosmos. Harvard, 1932. Historical survey of theories of structure of the universe.

Tolman, R. C. The Present Status of Cosmology. *Scientific Monthly*, *43*, 491–507, 1936; *44*, 20–40, 1937. Richly illustrated account of growth of theory and observational evidence.

II. *Original Papers and Technical Reports.*

Calinon, A. Les Espaces Géométriques I. *Rev. Philosophique*, *27*, 588–595, 1889. Proposes space of curvature which varies with time.

Einstein, A. Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. *S.-B. preuss. Akad. Wiss.*, pp. 142–152, 1917. Derivation of finite static universe on general theory of relativity.

Friedmann, A. Über die Krümmung des Raumes. *Zeits. f. Phys.*, *10*, 377–386, 1922. Proposes expanding universe as alternative to those of Einstein and De Sitter.

v. Helmholtz, H. Schriften zur Erkenntnistheorie, II. Über die Tatsachen, die der Geometrie zugrunde liegen. Springer, Berlin, 1921. Reprint of v. Helmholtz's 1868 *Göttingen Nachrichten* answer to Riemann's *Habilitationsrede*.

Hubble, E. The Luminosity Function of Nebulae II, The Luminosity Function as Indicated by Residuals in Velocity-Magnitude Relations. *Astrophys. Jour.*, *84*, 301–326, 1936. Definitive formulation of velocity-magnitude relation.

—— Effects of Red Shifts on the Distribution of Nebulae. *Astrophys. Jour.*, *84*, 445–482, 1936. Number-magnitude relation in five surveys out to *m* 21, and interpretation.

—— and Tolman, R. C. Two Methods of Investigating the Nature of the Nebular Red-Shift. *Astrophys. Jour.*, *82*, 391–426, 1935. Theoretical treatment of nebular diameter-magnitude and number-magnitude relations.

Klein, F. Vorlesungen über nicht-euklidische Geometrie. Springer, Berlin, 1928. Systematic treatise on non-Euclidean geometry.

Lemaître, G. Un Univers Homogène de Masse Constante et de Rayon Croissant, Rendant Comte de la Vitesse Radiale de Nébuleuses Extra-Galactiques, *Ann. Soc. Sci. Bruxelles*, *47A*, 49–59, 1927. Shows that Einstein universe of increasing radius can account for recession of nebulae.

Manning, H. P. Non-Euclidean Geometry. Ginn, New York, 1901. An elementary text.

McVittie, G. C. Cosmological Theory. Methuen, London, 1937. Concise résumé of relativistic theory and Milne's alternative.

Milne, E. A. Relativity, Gravitation, and World-Structure. Oxford, 1935. Proposes model based on postulates other than general relativity.

Riemann, B. Über die Hypothesen, welche der Geometrie zu Grunde liegen.

Springer, 1919. Reprint, edited (with notes) by H. Weyl, of Riemann's 1854 *Habilitationsrede.*

ROBERTSON, H. P. On the Foundations of Relativistic Cosmology, *Proc. Nat. Acad. Sci.*, *15*, 822–828, 1929. Derivation of suitable space-time geometry on general theory of relativity.

—— Relativistic Cosmology. *Rev. Mod. Phys.*, *5*, 62–90, 1933. Systematic report on theory, with extensive bibliography.

—— Kinematics and World-Structure. *Astrophys. Jour.*, *82*, 284–301, 1935. Examines relationship between any attack, such as Milne's, and relativity solution.

RUSSELL, B. The Foundations of Geometry. Cambridge, 1897. Reëxamination of Kant's position; a priori and empirical elements in geometry.

SCHWARZSCHILD, K. Ueber das zulässige Krummungsmaass des Raumes. *Vierteljahrsschrift der Astronomischen Gesellschaft*, *35*, 337–347, 1900. Determination of upper limit to curvature on basis of observational data available in 1900.

SHAPLEY, H. Note on the Problem of the Expanding Universe. *Proc. Nat. Acad. Sci.*, *24*, 148–154, 1938. Bearing of density-gradients on interpretation of number-magnitude relation.

DE SITTER, W. On the Relativity of Inertia. Remarks Concerning Einstein's latest Hypothesis. *Proc. Akad. Wetensch. Amsterdam*, *19*, 1217–1225, 1917. Original treatment of De Sitter universe.

TOLMAN, R. C. The Effect of the Annihilation of Matter on the Wave Length of Light from the Nebulae. *Proc. Nat. Acad. Sci.*, *16*, 320–337, 1930. Derivation of geometry from general relativity, and physical consequences.

—— Relativity, Thermodynamics, and Cosmology. Oxford, 1934. Complete discussion of relativistic models in chap. X.

WEYL, H. Zur allgemeinen Relativitätstheorie. *Phys. Zeits.*, *24*, 230–232, 1923. Deduces linear velocity-distance relation in De Sitter universe.

It is a pleasure to acknowledge my indebtedness to Dr. Edwin P. Hubble, of the Mount Wilson Observatory, Carnegie Institution of Washington, for allowing me to reproduce here his chart illustrating the velocity-distance relation; and to the Huntington Library, San Marino, California, for permission to reproduce here the Digges' diagram from a book in their collections.

CHAPTER VII

BLACKETT, P. M. S. The Cosmic Radiation. Hermann & Cie., Paris, 1935.

—— The Cosmic Radiation, *Roy. Soc. Arts.*, *85*, 893, 1937.

Buettner, K. The Penetrating Radiation. Handbuch der exper. Physik, XXV_1 (Wien-Harms), 479–523, 1928.

Corlin, A. Introductory Chapter to Cosmic Rays in Northern Sweden. *Ann. Obs. Lund.*, No. 4, 1934.

Darrow, K. K. Cosmic Rays—What Physicists Have Learned About Them. *Elect. Eng.*, 221–228, April, 1933.

Johnson, T. H. *Rev. Mod. Phys.*, *10*, 193, 1938.

Kolhorster, W., and Tuwin, L. Physical Problems of the Cosmic Radiation. Leipzig, 1934, 302 pp.

Leprince-Ringuet, L. The Cosmic Rays. Hermann & Cie., Paris, 1934.

Millikan, R. A. Cosmic Rays, Macmillan, 1939.

Millikan, R. A. On the Cosmic Rays. *Ann. de l'Inst. H. Poincaré*, 447–464, 1932.

Rossi, B. Cosmic Rays. Hermann & Cie., Paris, 1935.

Steinke, E. G. The Cosmic Radiation, *Ergeb. d. Exakt. Naturwiss.*, *13*, 89–134, 1934; Handbuch der Physik, $XXIII_2$, 521, 1933.

Symposium on Cosmic Rays. *Rev. Mod. Phys.*, *11*, 121, 1939.

A very extensive bibliography is contained in the references given in the publications listed above.

CHAPTER VIII

Abramson, H. A. Modification of the Northrop-Kunitz microcataphoresis cell. *Jour. Gen. Physiol.*, *12*, 469, 1928–29.

Brown, A. S., and MacInnes, D. A. The determination of activity coefficients from the potentials of concentration cells with transference. I, Sodium chloride at 25°. *Jour. Am. Chem. Soc.*, 57, 1356, 1935.

Burton, E. F. Properties of electrically prepared colloidal solutions. *Phil. Mag.*, (6) *11*, 425, 1906.

Cady, H. P., and Longsworth, L. G. A modification of the moving boundary method for the determination of transference numbers. *Jour. Am. Chem. Soc.*, *51*, 1656, 1929.

Debye, P., and Hückel, E. Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen. *Physik. Z.*, *24*, 185, 1923.

Ellis, R. Die Eigenschaften der Ölemulsionen. I. Die elektrische Ladung. *Z. physik. Chem.*, *78*, 321, 1912.

Franklin, E. C., and Cady, H. P. On the velocities of the ions in liquid ammonia solutions. *Jour. Am. Chem. Soc.*, *26*, 499, 1904.

Kohlrausch, F. Über Konzentrationsverschiebungen durch Elektrolyse im Inneren von Lösungen und Lösungsgemischen. *Ann. Physik*, *62*, 209, 1897.

LAMM, O. Measurements of concentration gradients in sedimentation and diffusion by refraction methods. Solubility properties of potato starch. *Nova Acta Regiae Soc. Sci. Upsaliensis* (No. 6), *10*, 1937.

LANDSTEINER, K., LONGSWORTH, L. G., and VAN DER SCHEER, J. Electrophoresis experiments with egg albumins and hemoglobins. *Science*, *88*, 83, 1938.

LAUE, M. VON. Die Wanderung von Unstetigkeiten in elektrolytischen Lösungen. *Z. anorg. Chem.*, *93*, 329, 1915.

LODGE, O. On the migration of ions and on experimental determination of absolute ionic velocity. Report of the British Association for the Advancement of Science, Birmingham, pp. 389–412, 1886.

LONGSWORTH, L. G. (1) Transference numbers of aqueous solutions of some electrolytes at 25° by the moving boundary method. *Jour. Am. Chem. Soc.*, *57*, 1185, 1935.

—— (2) A modification of the schlieren method for use in electrophoretic analysis. *Jour. Am. Chem. Soc.*, *61*, 529, 1939.

—— and MACINNES, D. A. (1) Transference numbers of lanthanum chloride at 25° by the moving-boundary method. *Jour. Am. Chem. Soc.*, *60*, 3070, 1938.

—— (2) An electrophoretic study of nephrotic sera and urine. *Jour. Exp. Med.*, *71*, 77, 1940.

—— SHEDLOVSKY, T., and MACINNES, D. A. Electrophoretic patterns of normal and pathological human blood serum and plasma. *Jour. Exp. Med.*, *70*, 399, 1939.

MACINNES, D. A., and BRIGHTON, T. B. The moving-boundary method for determining transference numbers. III, A novel form of apparatus. *Jour. Am. Chem. Soc.*, *47*, 994, 1925.

MILLER, W. L. The theory of the direct method of determining transport numbers. *Z. physik. Chem.*, *69*, 436, 1909.

MILNER, S. R. (1) Virial of a mixture of ions. *Phil. Mag.*, (6) *23*, 551, 1912.

—— (2) The effect of interionic forces on the osmotic pressure of electrolytes. *Phil. Mag.*, (6) *25*, 742, 1913.

NORTHROP, J. H., and KUNITZ, M. An improved type of microscopic electro-cataphoresis cell. *Jour. Gen. Physiol.*, *7*, 729, 1924–25.

ONSAGER, L. (1) Zur Theorie der Elektrolyte, I. *Physik. Z.*, *27*, 388, 1926.

—— (2) Zur Theorie der Elektrolyte, II. *Physik. Z.*, *28*, 277, 1927.

PHILPOT, J. ST. L. Direct photography of ultracentrifuge sedimentation curves. *Nature*, *141*, 283, 1938.

SHEDLOVSKY, T., and MACINNES, D. A. (1) The determination of activity coefficients from the potentials of concentration cells with transference. II, Hydrochloric acid at 25°. *Jour. Am. Chem. Soc.*, *58*, 1970, 1936.

SHEDLOVSKY, T., and MACINNES, D. A. (2) The determination of activity coefficients from the potentials of concentration cells with transference. III, Potassium chloride. IV, Calcium chloride. *Jour. Am. Chem. Soc.*, *59*, 503, 1937.
—— (3) The determination of activity coefficients from the potentials of contration cells with transference. V, Lanthanum chloride at 25°. *Jour. Am. Chem. Soc.*, *61*, 200, 1939.
STENHAGEN, E. Electrophoresis of human blood plasma. Electrophoretic properties of fibrinogen. *Biochem. Jour.*, *32*, 714, 1938.
SVENSSON, H. Direkte photographische Aufnahme von Elektrophorese-Diagrammen. *Kolloid-Z.*, *87*, 181, 1939.
TISELIUS, A. A new apparatus for electrophoretic analysis of colloidal mixtures. *Trans. Faraday Soc.*, *33*, 524, 1937.
WEBER, H. (1) Über die Differentialgleichungen der elektrolytischen Verschiebungen. *Sitzungsber. k. preuss. Akad. Wissensch. Berlin*, 936, 1897.
—— (2) Die partiellen Differential-Gleichungen der mathematischen Physik. 5th ed. Braunschweig, *1*, 195–203, 503–527, 1910.

CHAPTER IX

ARCHIBALD, W. J. The Process of Diffusion in a Centrifugal Field of Force. *Phys. Rev.*, *53*, 746; *54*, 371, 1938.
ASTON, F. W. Mass Spectra and Isotopes. Longmans, Green, 1933.
BAUER, J. H., and PICKELS, E. G. Improved Air-driven Type Ultracentrifuge for Molecular Sedimentation. *Jour. Exp. Med.*, *65*, 565, 1937.
BEAMS, H. W., and KING, R. L. Survival of Ascaris Eggs After Centrifuging. *Science*, *84*, 138, 1936.
BEAMS, J. W. An apparatus for obtaining High Speeds of Rotation. *Rev. Sci. Inst.*, *1*, 667, 1930.
—— High Speed Centrifuging. *Rev. Mod. Phys.*, *10*, 245, 1938. See for references.
—— and PICKELS, E. G. The Production of High Rotational Speeds. *Rev. Sci. Inst.*, *6*, 299, 1935.
—— and SKARSTROM, C. The Concentration of Isotopes by the Evaporative Centrifuge Method. *Phys. Rev.*, *56*, 266, 1939.
HARVEY, E. N. The Microscope Centrifuge and Some of its Applications. *Jour. Frank. Inst.*, *214*, 1, 1932.
HENRIOT, E., and HUGUENARD. Sur la réalisation de très grands vitesses de rotation. *Compt. Rend.*, *180*, 1389, 1925.
MULLIKEN, R. S. The Separation of Isotopes by Thermal and Pressure Diffusion. *Jour. Am. Chem. Soc.*, *44*, 1033, 1922.

MASON, M., and WEAVER, W. The Settling of Small Particles in a Fluid. *Phys. Rev.*, *23*, 412, 1924.

NICHOLS, J. B. Application of the Ultracentrifuge to Some Colloid Physical Problems. *Physics*, *1*, 254, 1931.

SVEDBERG, T. The Ultracentrifuge and the Study of High-Molecular Compounds. *Nature*, *139*, 1051, 1937.

—— and PEDERSEN, K. O., in coöperation with Bauer, J. H.; Pickels, E. G.; Boestad, G.; Kraemer, E. O.; Nichols, J. B.; Lamm, O.; McFarlane, A. S.; Signer, R. The Ultracentrifuge. Oxford, 1940. See for references.

TOLMAN, R. The Electromotive Force Produced in Solutions by Centrifugal Action. *Jour. Am. Chem. Soc.*, *33*, 121, 1911.

UREY, H. C. Separation of Isotopes. Reports on Progress in Physics. The Physical Society, *6*, 48, 1940.

WYCKOFF, R. W. G. The Ultracentrifugal Study of Virus Proteins. *Proc. Am. Phil. Soc.*, *77*, 455, 1937.

—— The Ultracentrifugal Purification and Study of Macromolecular Proteins. *Science*, *86*, 92, 311, 1937.

CHAPTER X

Abbreviations.

A.T. Aerodynamic Theory, 6 volumes. Julius Springer, Berlin, 1935. Edited by W. F. Durand.

F.D. Modern Developments in Fluid Dynamics, 2 volumes. Oxford Univ. Press, 1938. Edited by S. Goldstein.

N.A.C.A. National Advisory Committee for Aeronautics, Washington, D.C.

R. and M. Reports and Memoranda, Aeronautical Research Committee, London.

Jour. A.S. Journal of the Aeronautical Sciences, New York.

Jour. R.A.S. Journal of the Royal Aeronautical Society, London.

1. Aircraft Year Book. Aeronautical Chamber of Commerce, New York.
2. Civil Aeronautics Act. Office of Public Documents, Washington, D.C.
3. GEER, W. C. An Analysis of the Problem of Ice on Airplanes, *Jour. A.S.*, September, 1939.
4. KEEN, R. Wireless Direction Finding, chap. XV, *The Wireless World* (London), 1938.
   Electronic Navigation Equipment, Pender Electrical Eng. Handbook, vol. 5, sec. 16, p. 33.
5. MAGOUN, F. A., and HODGINS, E. A History of Aircraft. McGraw-Hill, New York, 1931.

6. All the World's Aircraft, 1938 annual. Sampson Low, Marston Co., London.
7. LANCHESTER, F. W. Aerodynamics. Constable, London, 1907.
8. KUTTA, W. M. Auftriebskräfte in strömenden Flüssigkeiten. *Illustrierte Aeronautische Mitteilungen*, July, 1902.
9. JOUKOWSKI, N. Aerodynamique. Gauthier-Villars, Paris, 1916.
10. PRANDTL, L. Applications of Modern Hydrodynamics to Aeronautics. N.A.C.A. Tech. Report 116, 1921.
11. GIACOMELLI, R. Historical Sketch. A.T., vol. 1, div. D.
12. LORD RAYLEIGH. Application of Principle of Dynamical Similarity. R. and M., 15, 1909.
13. BRYAN, G. H. Stability in Aviation. Macmillan, London, 1911.
14. BAIRSTOW, L. Applied Aerodynamics. Longmans, Green, London, 1920.
15. DRZEWIECKI, S. Des Hélices Aériennes. Lib. of Aero., Paris, 1909.
16. EIFFEL, G. The Resistance of the Air and Aviation, trans. J. C. Hunsaker. Constable, London, 1913.
17. PRANDTL, L. Über Flüssigskeitsbewegung bei sehr kleiner Reibung. N.A.C.A. Tech. Memorandum 452, 1904.
18. EDWARDS, FRARY, JEFFRIES. The Aluminum Industry. McGraw-Hill, New York, 1930, II, 135, 233.
19. JUNKERS, H. Metal Airplane Construction. *Jour. R.A.S.*, vol. 27, September, 1923, p. 406.
20. The Fairey Aviation Co., Ltd., All the World's Aircraft, p. 87a, 1920 (also same, 1919).
21. HANDLEY-PAGE, F. The Handley-Page Wing. *Jour. R.A.S.*, June, 1921, and August, 1928.
22. FEDDEN, A. H. R. Radial Air-Cooled Engines. *Jour. R.A.S.*, April, 1924.
23. —— The Single Sleeve Valve. *Jour. S.A.E.*, September, 1938.
24. CURTISS, G. H., and POST, A. The Curtiss Aviation Book. Frederick A. Stokes Co., New York, 1912.
25. REED, S. A. Reed One-Piece Metal Propeller. *Aerial Age* (New York), April, 1923.
26. CALDWELL, F. W. Aircraft Propeller Development. *Jour. S.A.E.*, August, September, 1934.
27. —— Hydromatic Propeller. *Aviation*, July, 1938.
28. MIDGLEY, T., and BOYD, A. Chemical Control of Detonation. *Industrial and Engineering Chemistry*, October, 1922.
TAYLOR, C. F., and TAYLOR, E. S. The Internal Combustion Engine. International Text Book Co., Scranton, 1938.
29. MILLER, H. B. Shooting the Catapult. *U.S. Naval Institute Proc.*, Annapolis, April, 1933.

30. Airways and Airports. See chapter in reference 1.
LYON, T. C. Practical Air Navigation. U.S. Coast and Geodetic Survey, Special Pub. 197, 1935.
31. ESPENSCHEID, L., and NEWHOUSE, R. C. A Terrain Clearance Indicator. *Jour. A.S.*, February, 1939.
32. OCKER, W. C., and CRANE, C. J. Blind Flight in Theory and Practice. Naylor Printing Co., San Antonio, 1932.
JORDANOFF, A. Through the Overcast. Funk and Wagnalls, New York, 1938.
33. The Aluminum Industry. Reference 18, page 302.
34. Fleetwing's Seabird. *Aviation* (New York), October, 1936.
Stainless Steel Aircraft. *Aero Digest* (New York), July, 1938.
35. WARNER, E. P. Parasite Drag. Chap. XII, Airplane Design-Performance. McGraw-Hill, New York, 1936.
Cowling and Cooling. N.A.C.A. Tech. Reports 313, 314, 332, 436, 592, 593, 595, 662.
36. Airfoils. N.A.C.A. Tech. Reports 460, 610, 628.
37. Propeller Characteristics. N.A.C.A. Tech. Reports 350, 564, 594, 639, 640, 642, 643, 650, 658.
38. Military and Civil Airplanes. *Flight* (London), April 27, 1939.
39. HOOD, M. J. The Effect of Surface Irregularities on Wing Drag. N.A.C.A. Tech. Notes 695 and 724.
40. PRANDTL, L. The Mechanics of Viscous Fluids. A.T., vol. 3, div. G.
41. DURAND, W. F. Fluid Mechanics. A.T., vol. 1, div. B.
42. LAMB, H. Hydrodynamics. Chap. IV. 6th ed. Cambridge Univ., 1932.
43. Flowpast Asymmetrical Cylinders. F.D., vol. 2, chap. X.
44. WARNER, E. P. High-Lift Devices. Chap. X, reference 35.
45. WEICK, F. E. Aircraft Propeller Design. McGraw-Hill, New York, 1930, chap. VIII.
N.A.C.A. Tech. Report 639.
46. The Speed Record Raised. *Flight* (London), p. 1152, November 1, 1934.
47. Scooped Yet Again. *The Aeroplane* (London), *41*, 419, April 5, 1939.
48. WARNER, E. P. Compressibility. Chap. V, p. 111; chap. VII, p. 221, reference 35.
49. Compressibility Effects on Airfoils. N.A.C.A. Tech. Reports 463, 492, 646.

INDEX